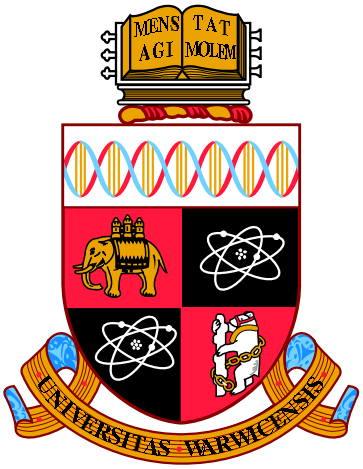


Engineering Stress Resilient Plants Using Gene Regulatory Network Rewiring

by

Iulia Gherman

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

School of Engineering

September 2018

Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to thank my supervisors, Katherine, David and Declan, for their support and guidance throughout my PhD. A huge thanks also to the Denby group for sharing in my woes (not knowing what to bake for cake club) but also in the good times - Elspeth, Maura, Sarah, Gill, Adam, Rachael, James, Krzysztof. You are all good eggs! Mathias, Chris - thank you for the great mentorship. Alex, thanks for the RNA troubleshooting; Ana & Silke thank you for the plasmids and advice! Thank you to my friends and family for all their love and support; I would not have been able to do it without you. Ed - you're awesome. Thanks for all the food and understanding.

Declarations

This thesis is presented in accordance with the regulations for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. All of the work in this thesis has been undertaken by myself except where otherwise stated.

Publications resulting from this thesis

Chapter 3 Foo, M.*, Gherman, I.*, Zhang, P., Bates, D. G., & Denby, K. (2018). A Framework for Engineering Stress Resilient Plants using Genetic Feedback Control and Regulatory Network Rewiring. *ACS Synthetic Biology*. *Joint first authors.

Chapter 4 Penfold C.A., Gherman I., Sybirna A., & Wild D.L. Inferring gene regulatory networks from multiple datasets. In *Gene Regulatory Networks* (pp. 251-282). Humana Press, New York, NY.

Foo, M., Gherman, I., Denby, K. J., & Bates, D. G. (2017). Control strategies for mitigating the effect of external perturbations on gene regulatory networks. *IFAC-PapersOnLine*, 50(1), 12647-12652.

Abstract

In spite of advances in food production brought on by the Green Revolution, the challenge of providing access to nutritious, safe food that has been grown sustainably is considerable. One such barrier to food security is biotic stress - infection with pathogens such as bacteria, fungi and oomycetes impact negatively on plant growth and survival. Synthetic biology, an interdisciplinary field combining biology, engineering and mathematics, is a promising tool for understanding and developing stress tolerant plants.

The response of the model plant *Arabidopsis thaliana* to biotic and abiotic stresses involves the transcriptional reprogramming of thousands of genes. Among these differentially expressed genes are transcription factors, which form complex causal networks specific to the stress in question. This thesis focuses on network rewiring as a tool for enhancing the Arabidopsis response to stress, in particular to *Botrytis cinerea* infection. This is a model system for studying plant-necrotrophic pathogen interactions and as such, a large amount of data are available, including a high-resolution transcriptomic time series of Arabidopsis during *B. cinerea* infection. This was used to construct gene regulatory networks with hundreds of transcription factors that are differentially expressed, in order to obtain a systems view of the effects of infection and the relationships between these regulators. Rewiring was applied to subnetworks of the original network using two different methodologies: control engineering, and Gaussian process dynamical systems. The former focuses on eliminating the effects of perturbation on a single node in a small 9-gene network, and requires detailed parameterisation of biological processes such as mRNA degradation and transcription rates. The latter provides a general modelling framework for optimising the overall expression of genes in a larger 70 gene subnetwork that eschews parameterisation or definition of a precise function for modelling relationships between genes.

The process of generating stably transformed and rewired Arabidopsis is long and requires growing hundreds of plants for each construct. In order to test the hypotheses generated by such computational tools quickly and on a large scale, Arabidopsis protoplasts treated with chitin were trialled as a model system for studying plant defence responses to *B. cinerea*. RNAseq analysis of protoplasts was used to determine the similarities and differences between the defence responses triggered in protoplasts and in Arabidopsis plants. Both protoplasts and plants were also rewired, and gene expression measurements used to understand the effects of this genetic engineering on the defence response of each.

Abbreviations

ABA	Abscisic Acid
AGI	Arabidopsis Gene Identifier
ARD	Automatic Relevance Determination
AtRegNet	*Arabidopsis thaliana* Regulatory Network
AUPR	Area Under the Precision-Recall curve
AUROC	Area under the Receiver-Operater Characteristic curve
B. cinerea	*Botrytis cinerea*
bHLH	Basic Helix Loop Helix
bp	Base Pair
CDS	Coding Sequence
ChIP-chip	Chromatin Immunoprecipitation and DNA Microarray
ChIP-seq	Chromatin Immunoprecipitation and DNA Sequencing
Col-0	Ecotype Columbia-0
CSIGREN	CSI and GRENITS model consensus
DAMP	Damage-Associated Molecular Pattern
DE	Differentially Expressed
DEG	Differentially Expressed Gene
DREAM	Dialogue for Reverse Engineering Assessments and Methods
E. coli	*Escherichia coli*
EIL	EIN3-like
EIN	Ethylene Insensitive
ER	Endoplasmic Reticulum
ERF	Ethylene Response Factor
ET	Ethylene
ETI	Effector-Triggered Immunity
FLS2	Flagellin-Sensing 2
G(1)	Gene 1
GENInfTIG	GENIE3, Inferelator and TIGRESS model consensus
GNW	GeneNetWeaver
GO	Gene Ontology
GP	Gaussian Process
GPDS	Gaussian Process Dynamical System
GPR	Gaussian Process Regression

GRN	Gene Regulatory Network
hpi	Hours Post Infection
HR	Hypersensitive Response
JA	Jasmonic Acid
KO	Knock-Out
LARS	Least Angle Regression
LB	Luria-Bertani
LRR-RK	Leucine-Rich Repeat Receptor Kinase
MAMP	Microbe-Associated Molecular Pattern
MCMC	Markov Chain Monte Carlo
miRNA	MicroRNA
mRNA	Messenger RNA
MSE	Mean Squared Error
NADPH	reduced Nicotinamide Adenine Dinucleotide Phosphate
NASC	Nottingham Arabidopsis Stock Centre
NB-LRR	Nucleotide-Binding Site Leucine-Rich Repeat
OE	Over-Expression
P. syringae	*Pseudomonas syringae*
PAMP	Pathogen-Associated Molecular Pattern
PPI	Protein-Protein Interaction
PR AUC	Precision-Recall Area Under the Curve
PRR	Pattern-Recognition Receptor
PTI	Pattern-Triggered Immunity
R gene	Resistance Gene
ROC AUC	Receiver-Operating Characteristic Area Under Curve
RNAseq	RNA sequencing
RT-qPCR	quantitative Reverse Transcription Polymerase Chain Reaction
ROS	Reactive Oxygen Species
SA	Salicylic Acid
SBML	Systems Biology Markup Language
SDE	Stochastic Differential Equation
SE	Squared Exponential
siRNA	Short Interfering RNA
tl-CLR	time-lagged Context Likelihood of Relatedness
TOFDE	Time OF Differential Expression
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
UTR	UnTranslated Region
WT	Wild-Type
Y1H	Yeast-1-Hybrid

Table 1: Abbreviations used in this work

Chapter 1

Introduction

1.1 Synthetic biology

Synthetic biology is an emerging discipline encompassing engineering principles applied to biological systems. Engineering principles of modularity, standardisation, abstraction, predictability and the design-build-test approach are often incorporated to design biological systems with a particular function. Biological components are treated as 'building blocks' or 'parts' and used to build up these new, artificial systems. These building blocks are mostly genes, but can be anything from DNA and peptides to proteins to whole cells. The synthetic biology vision can be summed up by the Richard Feynman quote 'what I cannot create I do not understand'. This approach of joining standardised modules in new ways can be used to study the initial system to better understand its components and how they interact; in addition, as mentioned, it can be used to create new systems for a specific purpose.

The synthetic biology framework lends itself to a large number of applications ranging from synthetic peptide barrels (Thomson *et al.*, 2014) to the introduction of nitrogen fixation in cereals (Geddes *et al.*, 2015) to synthetic organelles to the development of biofuels (Georgianna and Mayfield, 2012) to synthetic pattern formation (Basu *et al.*, 2005) to the creation of artificial microbial communities (Großkopf and Soyer, 2014).

Pioneering synthetic biology research was conducted in bacteria and involved the construction of genetic circuits such as the repressilator (Elowitz and Leibler, 2000) and toggle switch (Gardner *et al.*, 2000). This so-called first wave of synthetic biology focused on simple functional modules with precise behaviours. In order to produce such modules, each individual part such as promoters, ribosome-binding sites, mRNA/protein degradation, terminators, localisation tags, phosphorylation site, needed to be characterised and standardised. Characterisation involves quantitatively measuring part performance *in vivo* or *in vitro*. For instance, the characterisation of an arabinose-inducible promoter can be done by attaching the promoter to *GFP*, and measuring the fluorescence of GFP at different concentrations of arabinose. Standardisation refers to the robustness seen in most engineering devices: light bulbs, batteries or wires can be used in many different circuits, as long as they are connected correctly. Similarly, if biological parts are

standardised, they should function well in different cellular backgrounds and environmental conditions. The creation of open-access libraries of characterised parts helped these efforts: some of these include the iGEM repository at http://parts.igem.org/, the JBEI-ICE platform (Ham *et al.*, 2012) at https://acs-registry.jbei.org/, the SynBioHub (McLaughlin *et al.*, 2018) at https://synbiohub.org/, and the SBOL Stack platform (Madsen *et al.*, 2016).

As the technology matured, the second wave of synthetic biology expanded its scope to a systems level approach (Purnick and Weiss, 2009) in terms of more complex circuitry (Seelig *et al.*, 2006) and additional biological chassis such as viruses (Citorik *et al.*, 2014) and eukaryotes (mammalian cells: Lienert *et al.* (2014); plants: Liu and Stewart (2015); yeast: Siddiqui *et al.* (2012)) were introduced. The chassis refers to the organism or host or genetic/biochemical framework where synthetic modules can be plugged in and out and is a metaphor lifted from engineering.

Engineering biology is challenging. The inherent robustness of organisms selected by evolution ensures that minimal changes occur during perturbations (Kitano, 2004). This resistance to change can be detrimental to synthetic biology efforts, which aim to introduce new pathways or re-engineer the original host pathways. Crosstalk with host components, even when expressing non-native, inactive fluorescent proteins can have unpredictable effects on expression levels (Cardinale *et al.*, 2013). Noise and randomness, while serving an important function in biology, prevent the robust functioning of constructs in heterogeneous cell populations (Rao *et al.*, 2002). Synthetic DNA constructs in plasmids, especially highly expressing ones, provide an unnatural burden on the chassis (Shachrai *et al.*, 2010) and can prove toxic (Ow *et al.*, 2006). These synthetic devices usually rely on host machinery and resource allocation of finite cellular resource can lead to less-than-optimal functioning of devices and slow cell growth (Qian *et al.*, 2017).

As a result, designing the behaviour of synthetic systems is non-trivial and often requires many iterations of the design-build-test cycle. New strategies are necessary to enable even simple designs to function as desired, for instance by dynamic allocation of synthetic ribosomes (Darlington *et al.*, 2018) or using feedback control to reduce burden (Ceroni *et al.*, 2018) and provide a constant gene expression unaffected by variation in copy number and genomic position (Segall-Shapiro *et al.*, 2018).

1.1.1 Synthetic biology in plants

Plants are a natural chassis for synthetic biology. They can sense and respond to their environment, are easy and cheap to grow, and are able to grow in a wide variety of conditions. Plants also produce a vast array of chemically-diverse secondary metabolites which are widely used in pharmaceuticals, dyes, flavours, insecticides and fragrances (Verpoorte and Alfermann, 2013).

Synthetic biology work being done in food crops usually involves enhancing crop yield and quality. This includes genetically engineering non-leguminous crops to develop root nodules, enhance the root microbiome, and insertion of enzymes for nitrogen fixation directly into the plant to reduce the amount of nitrogen fertilisers while sustaining yields

(Geddes *et al.*, 2015; Sreevidya *et al.*, 2006). Another considerable challenge with a large pay-off being tackled involves generating C3 crops such as wheat, rice, soybean and potato capable of C4 photosynthesis to enhance the efficiency of CO_2 assimilation and yield (Häusler *et al.*, 2002).

A second application is producing recombinant proteins for therapeutic purposes, industrial or biopolymer use in plants. They allow the synthesis of complex compounds which is too expensive and impractical by chemical means (such as precise enantiomers), in a more environmentally friendly production process, without any bacterial toxins or human pathogenic microbes that may exist in animal cultures or production systems (Boehm, 2007). Some compounds produced in plants include monoclonal antibodies for the West Nile Virus expressed in *Nicotiana benthamiana* (Lai *et al.*, 2010), spider silk proteins in Arabidopsis or *N. benthamiana* Hauptmann *et al.* (2013), biofuel from algae (Georgianna and Mayfield, 2012) and alkaloids for medicinal purposes (Glenn *et al.*, 2013).

Biosensing is another common application, as plants already use sophisticated signal transduction systems consisting of receptors, histidine kinases, and response regulators to sense and respond to their environment. Plants have been engineered as environmental monitors for pollutants, explosives or harmful chemicals such as TNT in the soil (Antunes *et al.*, 2011). Arabidopsis and tobacco have also been used as biosensors for pathogens such as *Pseudomonas syringae*, providing an early warning system for farmers (Fethe *et al.*, 2014).

Arabidopsis thaliana, thale cress, is a model plant organism due to its small, diploid genome and rapid - in comparison to other plants - life cycle of 5 to 7 weeks from seed to seed. It is very amenable to *Agrobacterium tumefaciens*-mediated genetic transformation. In addition, it produces a large number of seeds - around 6000 per plant, on average (Choe *et al.*, 2001) and requires minimal growth space. It has played an important role in advancing fundamental biology and was the first plant to have its genome fully sequenced by 2000 by the Arabidopsis Genome Initiative (Tabata *et al.*, 2000). The vast amount of knowledge gathered on the functioning of its genes and the tools available for work in Arabidopsis make it the perfect choice for genetic engineering. Techniques first developed in Arabidopsis can then be applied to other crops. Arabidopsis is the organism of choice for this thesis.

1.2 Plant stress and food security

Food insecurity has been threatening mankind for millenia, and is still a pressing issue, despite the improvements brought on by the Green Revolution (Evenson and Gollin, 2003). It exists when people do not have access to safe and nutritious food for social, economic or physical reasons. With the global population on track to reach 9.8 billion by 2050 (according to a 2017 UN report, 'World Population Prospects: The 2017 Revision'), providing access to enough safe and nutritious food in a sustainable way becomes even more of a challenge. The role of plants in tackling this challenge cannot be overstated, as they are both a direct food source, and used for the rearing

of livestock. The affluence of the average person is rising, therefore both the amount of food, and providing a wider variety are important considerations. As the affluence of a population and urbanisation increases, so does the demand for meat and dairy products. More than 30% of the cereal produced globally is diverted to feed livestock, which is energy inefficient - every kilocalorie of meat generated requires on average an input of 7 kilocalories of cereal feed (Tscharntke *et al.*, 2012). This is unsustainable for another reason - agriculture contributed 11% to the total greenhouse gas emission in 2010, primarily due to livestock and fertiliser use (Tubiello *et al.*, 2015).

A number of challenges stand in the way of global food security. Approximately a third of the food produced is thrown away. The reasons for this are different in the developing and developed worlds. Lack of storage technology, and insufficient infrastructure in the developing world leads to loss due to spoilage after harvest. Pre-retail losses are smaller in the developed world, but cheap prices, high cosmetic standards, and over-reliance on use-by dates leads to waste at commercial retailers and homes (Godfray *et al.*, 2010).

Food security relies on *access* to food rather than just sufficient crop production. For the last few decades, global production was sufficient to feed the global population, yet between 2010-2012, 870 million people remained hungry (FAO and WFP, 2012). Given that the solution to food insecurity seems to relate more to infrastructure, accessibility and the way food is used, rather than food production, investment in plant science and crop improvement has been falling as a response (Beintema *et al.*, 2009). However, population and consumption trends suggest that this is ill-advised. With a predicted population growth of 35% by 2050, this will necessitate the production of 85% more primary foodstuff in 2050 relative to 2013 levels (Long *et al.*, 2015). While average yields are on the rise, the projected trends are not nearly enough to satisfy future demand (Figure 1.1).

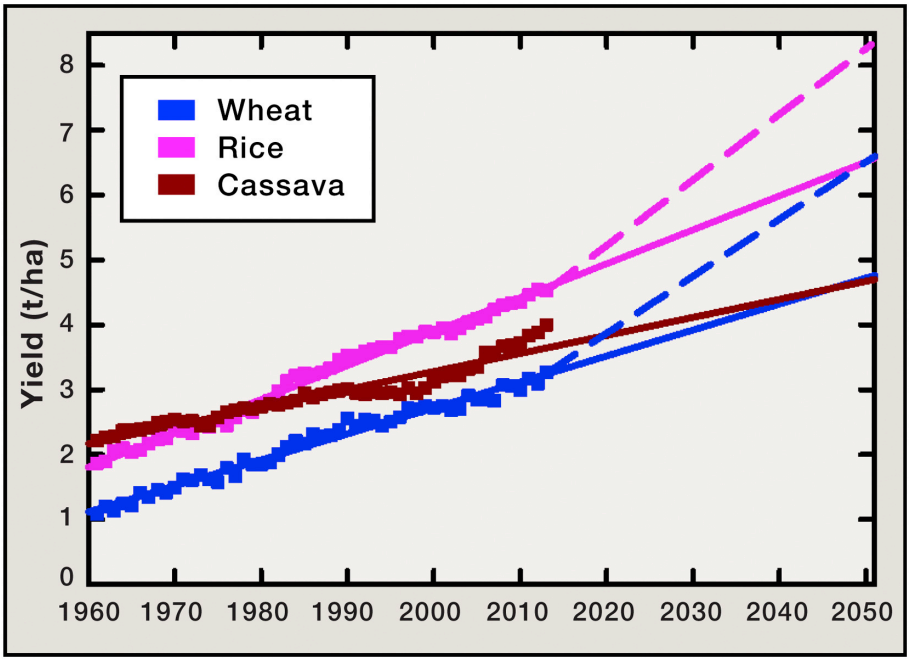


Figure 1.1: Average annual yields of wheat, rice and cassava in metric dry tons per hectare up to 2013. Solid lines are least-square regressions fitted to the data and extrapolated forward to 2050. Broken lines indicate the predicted demand for these crops. Image from Long *et al.* (2015).

Additionally, climate change is increasing growing season temperatures, bringing new problems. The summer average temperature at the end of the 21st century is predicted to exceed the hottest temperature on record in the tropics and subtropics (Battisti and Naylor, 2009). The 2003 heatwave in Europe was estimated to cause a loss of 30% in gross primary productivity (Ciais *et al.*, 2005). A significant proportion of agricultural land is saline - 22%, with global warming expanding this area and the amount of land in arid and semi-arid regions through a decrease in precipitation and enhanced evaporation (Wood *et al.*, 2000).

The contribution of agriculture to global warming (Tubiello *et al.*, 2015), the unsustainable use of groundwater for irrigation (FAO, 2011), groundwater, marine (Diaz and Rosenberg, 2008) and riverine (Turner *et al.*, 2003) pollution through fertiliser use and the loss of ecosystems (Gordon *et al.* (2010), McKenzie and Williams (2015)) highlights the need for a *sustainable* increase in productivity.

Therefore an increase in production efficiency, in a sustainable manner with regards to natural resources, is crucial for coping with the increasing global population and changing environment. The need for increased production should be met through higher yields achieved through efficient management practices rather than an increase in land (Tilman *et al.*, 2011). One way of achieving this is by producing stress-tolerant crops (Godfray *et al.*, 2010; Schmidhuber and Tubiello, 2007).

Stresses are changes in the environment which disrupt the ideal conditions for maintaining homeostasis and necessitate a response known as acclimation. As sessile organisms, plants are constantly exposed to biotic stresses (damage by other living organisms, such as infections from viruses, bacteria, fungi, oomycetes, loss of resources to parasites, and also mechanical damage from weeds and animals) and abiotic stresses (damage from environmental factors such as drought, flooding, high salinity, extreme temperatures, high light) in their environment that they cannot evade. These stresses produce a number of morphological, physiological, biochemical and molecular changes in plants, either as a direct effect or due to the defence mechanisms of the plant, which reduce plant growth and yields. Stresses are responsible for reducing productivity by 65-87%, depending on the crop (Buchanan *et al.*, 2015).

Abiotic stress, such as drought, non-optimal temperatures, strong light, salinity and poor soil nutrition, has the largest impact on crop yields, with an average loss of 50% for most major crop plants in a year (Boyer (1982), Bray (2000)). Plants are frequently exposed to a combination of abiotic stresses - for instance, drought conditions are associated with increased salinity, and heat (Mittler, 2006). This is a problem because the combination of individual stresses, such as heat and drought stress on maize, barley, sorghum and different grasses has a higher impact (Craufurd and Peacock, 1993; Heyne *et al.*, 1940; Jiang and Huang, 2001; Savin and Nicolas, 1996; Wang and Huang, 2004).

At least 10% of crops are lost to infections from pathogens such as viruses, bacteria, Oomycetes, fungi, nematodes and parasitic plants (Strange and Scott, 2005). 50% of barley and in excess of 80% of sugar beet and cotton can be lost due to pest and pathogens, if biocides are not in place, while actual losses stand at 26-30% for sugar beet, barley, soybean and cotton, and 35%, 39% and 40% for maize, potatoes and rice, respectively (Oerke and Dehne, 2004). Fungi and oomycetes alone decimate enough major crops to feed 0.6 billion people (Fisher *et al.*, 2012). A loss of 10% in rice amounts to enough rice to feed 60 million people for a year - correspondingly, even small reductions in crops lost to stress can benefit millions.

The dependence of a large proportion of the population on a single or a few crop species can lead to disaster should new pathogens arise to which the crops are susceptible. The Great Bengal Famine in 1943 and the potato blight of the 1840s are two such epidemics that cost millions their lives. As agriculture becomes increasingly globalised, crops are grown from a narrower genetic base, and may be far from their place of origin. This makes it harder to develop resistance against evolving pathogens in the area they are grown in. If said newly-evolved pathogens should spread, the plants would be especially vulnerable. Pathogens are also encouraged to spread by the increase in mean global temperature - a study of 612 crop pests found that they move polewards at a rate of 27 km per decade, on average (Bebber *et al.*, 2013).

A proposed solution for combating individual or combined stresses and enhancing tolerance is to create transgenic plants by manipulating protein-coding regions that are linked to tolerance (Cushman and Bohnert, 2000). A common way of discovering genes with phenotypes is quantitative trait loci (QTLs) analysis, which identifies sections of DNA responsible for phenotypes (reviewed in Bhatnagar-Mathur *et al.* (2008)). This ge-

netic engineering approach to enhancing stress tolerance eschews the usual slow rates of evolution but depends greatly on our knowledge of plant immunity and the plant stress response, and the availability of the beneficial alleles. Beneficial protein-coding regions (genes) can be inserted into the plant genome through molecular biology or conventional selective breeding, while genes detrimental to plant growth can be silenced. Turning single genes on and off is the simplest way of engineering crops; but it is possible to modify the timing, tissue-specificity, and expression level of proteins for a more specific response. A benefit of engineering single genes is that very specific targeted changes can be achieved. For instance, plants that lack osmoprotectants for dealing with osmotic stress can be made to express osmolytes such as glycine-betaine, proline and sugar alcohols (for example, mannitol, trehalose, myo-inositol and sorbitol) (Bhatnagar-Mathur *et al.*, 2008). These osmolytes protect protein complexes and the membrane, aiding their osmotic adjustment during times of drought, salinity and high temperature. Proline, for instance, achieves this by stabilising membranes and proteins, regulates redox potentials and scavenging free radicals (Hayat *et al.*, 2012). Another target for engineering are reactive oxygen species created during stresses: transgenic improvements to the detoxification strategy inbuilt in plants have enhanced tolerance to various stresses (Oberschall *et al.*, 2000; Roxas *et al.*, 1997). Synthetic biology tools and principles can also be used as an approach to tackling biotic stresses and enhance the plant stress response, and is the main topic of this thesis.

1.3 Molecular details of the Arabidopsis biotic stress response

The plant stress response starts with the perception of environmental signals through pattern recognition and conversion of these signals to a response appropriate for that particular stress in order to repel the pathogen (Mittler, 2006). The mechanisms of plant innate immunity are comparable to innate immunity in animals. However, unlike animals, plants do not have adaptive immunity or a specialised task force of immune cells dedicated to this purpose; every plant cell is capable of recognising pathogen molecules and effectors. Plants also rely on intercellular signals from infection sites to build up the immune response.

The plant responds in a myriad of ways to pathogens, including producing antimicrobial compounds and reinforcing the cell wall with additional polymers. Phytoalexins and defensins are some of the antimicrobials synthesised upon pathogen detection (Mazid *et al.*, 2011). Plants deposit callose and lignin at sites of pathogen penetration to increase the barrier to the pathogen; they can also deposit phytoalexins, chitinases, proteases and phenolics (Hückelhoven, 2007). The components of the various pathways were usually discovered through simple genetic screens where genes were knocked out and the impact on the pathway assessed. A breakdown of each step in the process, from sensing to responding to a pathogen threat, is provided in the following sections

1.3.1 Biotic stress perception

The defence response to a particular biotic stress can be triggered by the perception of one of three molecular danger signals: microbe-associated molecular patterns (MAMPs, which also used to be known as pathogen-associated molecular patterns, or PAMPs), damage-associated molecular patterns (DAMPs) and effectors (Jones and Dangl, 2006). MAMPs are a class of molecules absent from plants but essential and highly conserved in pathogens, such as chitin from the cell walls of fungi, β-glucans from oomycetes, and bacterial peptidoglycans, lipopolysaccharides, flagellin and elongation factor EF-Tu (Boller and He, 2009; Newman *et al.*, 2013). DAMPs are produced by the plant itself in response to damage caused by microbes; these can be plant elicitor peptides and oligogalacturonides from degradation of the cell wall (Choi and Klessig, 2016). Effectors, or virulence factors, are molecules produced by plant pathogens in order to alter host-cell function and structure and enhance pathogen fitness (Boller and Felix, 2009).

The defence response is triggered by pattern recognition receptors (PRRs) activated by MAMPs and DAMPs. One such MAMP is a conserved region of 22 amino acids in flagellin, the building block of bacterial flagella, which is vital for its pathogenicity (Naito *et al.*, 2008). The defence triggered by flagellin binding to PRRs is the core basal immune response, analogous to innate immunity in animals, and is called the PTI - PAMP-triggered immunity. However, microbes have learned to anticipate the PTI and respond by secreting molecules called effectors in order to suppress it. This has lead to the development of a secondary innate immunity known as effector-triggered immunity (ETI). This defence is based on intracellular recognition of pathogen effectors by specific nucleotide-binding leucine-rich repeat (NB-LRR) receptors, a type of plant resistance (R) genes. According to the guard hypothesis, R proteins can also monitor effector targets, rather than detecting the effector directly, to deduce whether a pathogen is modifying the normal cellular state of the plant (Dangl and Jones, 2001). If triggered, R proteins can give rise to the hypersensitive response (HR), leading to localised programmed cell death and systemic defence signalling, a response highly effective against biotrophic pathogens (Dodds and Rathjen, 2010).

The ETI and PTI share immune responses such as the oxidative burst, hormonal changes, the hypersensitive response and transcriptional reprogramming, as well as signalling pathways. In signalling terms, the ETI is a more intense and prolonged version of the PTI response. This can be seen in the length and amplitude of the ROS burst, MAP kinase signalling, and the robustness afforded by the compensatory/redundant structure of the ETI signalling network (Torres *et al.*, 2006; Tsuda *et al.*, 2009; Underwood *et al.*, 2007). The PTI can be triggered by non-pathogenic microbes, due to the highly conserved nature of PAMPs, which may explain the cost-saving benefits of having a weaker immune response initially (Tsuda and Katagiri, 2010). The ETI, on the other hand, is triggered by recognition of specific effectors, and can result in the expression of R genes that are highly specific to the pathogen encountered, such as the R genes R genes *RPP4, RPP7*, and *RPP8* for the oomycete *H. parasitica* (Katagiri, 2004). This sequence of events is a result of an evolutionary gene-for-gene arms race between plants

and their pathogens (reviewed in Gassmann and Bhattacharjee (2012), see Figure 1.2).

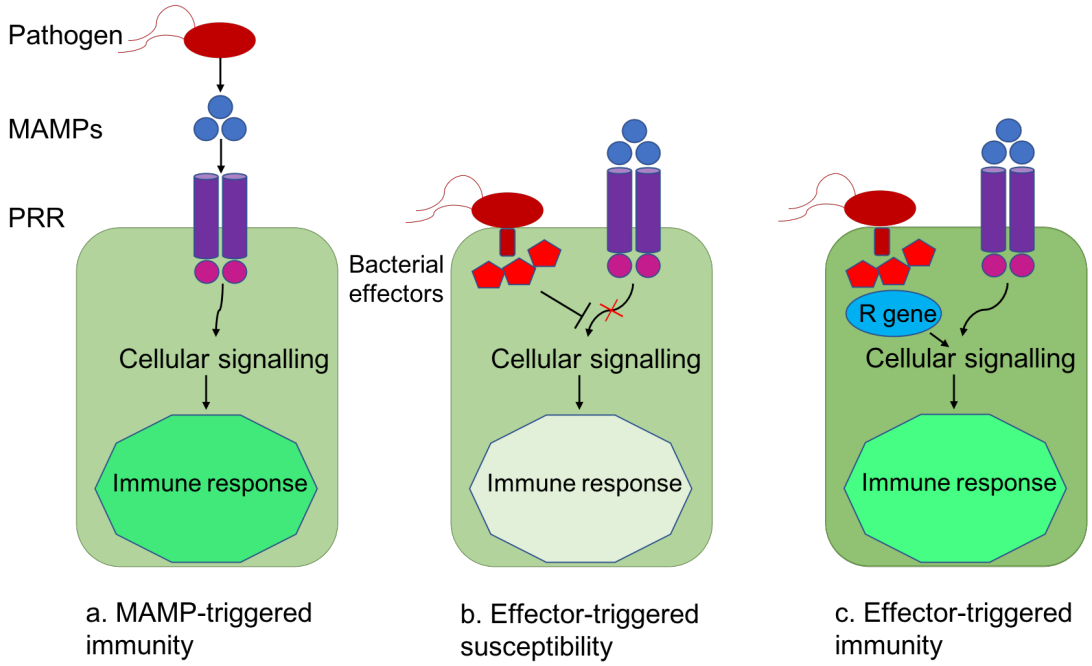


Figure 1.2: An overview of plant-pathogen interactions. (a) Pathogens are recognised by pattern recognition receptors (PRRs) based on their pathogen/microbe-associated molecular patterns (PAMPs/MAMPs), triggering a cascade of intracellular immune signalling, which activates the PAMP-triggered immunity (PTI). (b) Pathogens such as bacteria release effector proteins through type 3 secretion systems; this blocks the immune signalling and suppresses the immune response. For instance, *P. syringae* releases coronatine, HopI1 and AvrRpt2 to manipulate host signalling networks. (c) Plant resistance (R) proteins sense the effector proteins or their downstream effects and act to trigger the secondary immune response, the effector-triggered immunity (ETI). R genes such as RIN4, RPS2 and RPM1 modulate resistance to *P. syringae*. Figure based on Pieterse *et al.* (2009).

1.3.2 Signal transduction

Signal transduction is the mechanism through which pathogen attack is communicated to the cellular machinery that activates the defence response. The end goal of the complex signalling cascades is often the modulation of gene expression to increase the levels of proteins offering cellular protection, and the production of plant hormones such as salicylic acid (SA), abscisic acid (ABA), jasmonic acid (JA) and ethylene (ET) (Atkinson and Urwin, 2012; Bari and Jones, 2009). In addition, metabolic processes such as photosynthesis (Bilgin *et al.*, 2010), and secondary metabolism (the production of small molecules which are not essential for growth but aid in the plant's survival and

defence, for example flavonoids, terpenoids, alkaloids), as well as the response to abiotic stress such as drought, osmotic, cold, high light and oxidative stresses are differentially regulated (Windram *et al.*, 2012) in an effort to maximise resource usage to combat the pathogen threat.

Signals from PRRs and NB-LRRs are relayed through secondary messengers such as inositol phosphates and reactive oxygen species (ROS). Phosphorelay cascades consisting of phosphorylation (carried out by kinases) and dephosphorylation (carried out by dephosphorylases) events are set into motion by the ROS or a change in intracellular Ca^{2+} levels; these transduce and amplify the information signal (Xiong and Zhu, 2001). Mitogen-activated protein kinase (MAPK) signalling is ubiquitous in eukaryotes and generally consists of a MAPK kinase kinase (MAPKKK), which phosphorylates a MAPK kinase (MAPKK), which phosphorylates a MAPK. MAPK cascades can be triggered by both PTI and ETI (Dodds and Rathjen, 2010). For instance, the Flagellin Sensing 2 transmembrane protein responds to the highly conserved region of flagellin from *P. syringae*, and in conjuction with BRI1-Associated Kinase 1, triggers a MAPK cascade (Zipfel, 2008). This activates MAPKKs MKK4 and MKK5 to phosphorylate MAPKs MPK3 and MPK6, which then increase transcription of defence-related genes, including WRKY22 and WRKY29 (Asai *et al.*, 2002). Calcium-dependent protein kinases (CDPK) are also activated by MAMPs by sensing changing Ca^{2+} levels, and can be activated independently of the MAPK pathway. This was shown in CDPK mutants where the ROS burst was reduced and *P. syringae* growth was increased (Boudsocq *et al.*, 2010).

Far from being passive signal relays, over 80% NB-LRRs are predicted to localise to the nucleus after activation (Caplan *et al.*, 2008). From the nucleus they participate in the defence response by transcriptional reprogramming through their own transcription factor domains such as WRKY and DREF DNA-binding finger (BED) domains (Caplan *et al.*, 2008) or by associating with transcription factors, bypassing signal transduction cascades (Holt III *et al.*, 2002; Shen *et al.*, 2007).

NB-LRRs triggered by effectors or host cell effector targets result in ETI, which is an amplified version of the PTI, and HR, involving cell death. Cell death triggers SA-dependent signalling and vice versa through feedback loops, activating defence genes (Glazebrook, 2005). The HR requires the production of ROS in what is known as an oxidative burst. ROS have numerous functions: they contribute to the reinforcement of the cell wall, activate defence genes and signalling components, and generate phytoalexins and hydroxyl radicals which can directly kill pathogens (Torres, 2010). Reduced nicotinamide adenine dinucleotide phosphate (NADPH) oxidases AtrbohD and AtrbohF are responsible for most ROS production and mutants deficient in these have a less effective HR (Torres, 2010). The HR is most effective against biotrophic pathogens that need to grow on living host tissue, depriving them of their food source. For instance, the HR lowers water potential in *Pseudomonas syringae*, leading to water stress of the pathogen (Wright and Beattie, 2004). However, the HR response is detrimental to the plant in the case of infection by necrotrophs, pathogens that obtain nutrition by killing host tissue.

1.3.3 Phytohormones

Phytohormones are significant downstream targets of signalling cascades triggered by MAMPs or DAMPs. Plants produce a large variety of these phytohormones, including auxins, abscisic acid (ABA), salicylic acid (SA), ethylene (ET), jasmonates (JA), gibberellins (GA), cytokinins (CK), and brassinosteroids (BR) and these are involved in growth and developmental processes as well as the response to stress. SA, ET, JA and ABA signalling is heavily involved in the biotic stress response. The nature of the pathogen determines the response of the phytohormones - SA signalling is associated with defence against biotrophs and JA/ET signalling with defence against necrotrophs and herbivorous insects (Glazebrook, 2005). ABA was mainly thought to coordinate the abiotic stress response, however its involvement in the pathogen response is now established (Lee and Luan, 2012). There is much cross-talk between the signalling pathways, with the SA and JA/ET responses being mutually antagonistic (Bari and Jones, 2009). ABA also appears to suppress the JA/ET response (Lee and Luan, 2012). However, there is also evidence for coordination or coregulation between the pathways in terms of similar gene induction after hormone treatment (Schenk *et al.*, 2000). The sophisticated cross-talk of the different hormone pathways ensure that the plant response is tailored to the stress in question and unnecessary action is avoided. The most important hormones and their downstream effects are discussed individually below.

1.3.3.1 Salicylic acid

SA is a type of phenolic acid that is synthesised upon detection of pathogens at the site of infection. It activates the defence response to biotrophic and hemi-biotrophic pathogens, leading to the HR, and establishes systemic acquired resistance (SAR). SAR is long-term resistance to a pathogen developed by tissue away from the site of infection, and characterised by the expression of pathogenesis-related (PR) genes. It can be initiated by exposure to avirulent strains and be effective against the pathogenic strain, or even be induced by a pathogen and confer resistance to a broad host of viral, bacterial and fungal diseases (Durrant and Dong, 2004). Treatment of plants with SA can induce SAR.

NPR1 is the SA receptor which is deoligomerised upon binding to SA, and this active form interacts with TGA (Loake and Grant, 2007) and WRKY transcription factors (TFs) (Wang *et al.*, 2006) to mediate the defence response. It also modulates crosstalk with the JA pathway by suppressing LOX2, PDF1.2 and VSP proteins - all JA-responsive genes (Spoel *et al.*, 2003).

1.3.3.2 Jasmonic acid

JA is a lipid-derived compound which plays a role in numerous processes, from germination, to senescence, fruit ripening, stomatal opening, and tuber formation. It is also the main phytohormone implicated in the response to herbivores and necrotrophic pathogens. JA concentration increases at sites of infection and tissue damage. Defi-

ciencies in the JA-response pathway leads to decreased resistance to fungal pathogens (Thomma *et al.*, 1998).

The main components of JA signalling are JAR1, JIN1/MYC2, COI1, and Jasmonate ZIM-domain (JAZ) proteins. Under stress-free conditions, JAZ proteins normally repress JA signalling by binding to and repressing JA-responsive genes, such as the TF JIN1/MYC2 (Chini *et al.*, 2007), and other bHLH and MYB TFs that mediate JA-dependent anthocyanin accumulation and initiation of trichomes (Qi *et al.*, 2011). Upon stress or wounding, JAR1 conjugates isoleucine to JA, and this JA-isoleucine is perceived by the COI1 receptor. JA-isoleucine promotes the ubiquitination and subsequent degradation of JAZ proteins, inducing two distinct and antagonistic response pathways. The first allows JIN1/MYC2 to activate JA-responsive genes induced by wounding, such as VSP2, and repress defence genes, such as PDF1.2 and basic chitinases (Bari and Jones, 2009; Lorenzo *et al.*, 2004). COI1 removal of JAZ repression also activates ethylene response factor 1 (ERF1), ethylene insensitive 3 (EIN3) and EIN3-like 1 (EIL1), which repress the wounding pathway and activate pathogen-responsive genes (Lorenzo *et al.*, 2004; Zhu *et al.*, 2011). The pathways includes a negative feedback loop whereby JAZ proteins are induced by JA (Bari and Jones, 2009).

JA and ET signalling activate common genes, including PDF1.2 through the integrator ORA59 (Zarei *et al.*, 2011), and ERF1 through the second pathway mentioned above (Lorenzo *et al.*, 2003).

1.3.3.3 Ethylene

Ethylene is a biologically active gas and the simplest unsaturated hydrocarbon ($H_2C{=}CH_2$), and is synthesised from methionine, an amino acid (Bleecker and Kende, 2000). It is involved in a number of growth and development processes, such as senescence, and is responsible for climacteric fruit ripening (Lelièvre *et al.*, 1997). Production of ethylene in Arabidopsis is also enhanced by the presence of pathogens (Broekaert *et al.*, 2006). This leads to the production of cell wall-strengthening products such as hydroxyproline-rich proteins, PR genes with direct antimicrobial activity such as PR-2, PR-3 and defensins, and phytoalexins (Broekaert *et al.*, 2006). The major TFs downstream of ethylene are ERFs.

The ethylene receptors ETR1 and ERS1 are transmembrane proteins embedded in the endoplasmic reticulum (ER) membrane; both form homodimers with a hydrophobic pocket for ethylene binding (Chang *et al.*, 1993; Hua *et al.*, 1995). CTR1 is a kinase and repressor of EIN2 in normal circumstances, which associates with the ET membrane receptors in the ER. Binding of ET to the receptors causes a conformational change in CTR1 which releases EIN2 from its repression (Ju *et al.*, 2012). EIN2 is then free to activate EIN3 and EIL transcription factors (Ju *et al.*, 2012). The EIN3 TF binds its targets at the EIN3-binding site in the promoter of genes like ERF1, and ERF4, which then act downstream on genes like PDF1.2 and other defence response genes (Guo and Ecker, 2004).

1.3.3.4 Abscisic acid

ABA was thought to be responsible mainly for the response to environmental conditions such as cold, salinity, heat, drought and wounding as these increase ABA levels (Lata and Prasad, 2011). Its role in the abiotic stress response centres on production of hydrogen peroxide, leading to nitrogen monoxide-mediated stomatal closure (Bright *et al.*, 2006). Less is known about its molecular role in biotic stress. High ABA levels have been shown to increase susceptibility and low levels enhance resistance to biotrophic and necrotrophic pathogens including *P. syringae* pv *tomato*, *Peronospora parasitica* and *Botrytis cinerea* (Audenaert *et al.*, 2002; Mohr and Cahill, 2003; Thaler and Bostock, 2004). An ABA-dependent pathway encourages callose deposition as a physical barrier against the colonisation of necrotrophs *A. brassicicola* and *P. cucumerina* (Ton and Mauch-Mani, 2004).

It appears that ABA mediates the plant response by reducing the SA-mediated response (Thaler and Bostock, 2004) and inhibiting ET production (LeNoble *et al.*, 2004). One point of convergence is EIN2, which negatively regulates ABA biosynthesis (Ghassemian *et al.*, 2000). ABA also interacts synergistically and antagonistically with JA signalling (Mauch-Mani and Mauch, 2005). High ABA concentrations were found to reduce the levels of JA- and ET-responsive genes, even after the addition of methyl-JA or ET. This suggests that abiotic stress signalling takes precedence over biotic stress signalling (Anderson *et al.*, 2004). The TF MYC2/JIN1 is another point of convergence between ABA and JA pathways - upon JA signalling it activates the wounding response and insect defence and represses the response to pathogens, and it is also an activator of ABA signalling (Kazan and Manners, 2013).

1.3.4 Plant transcription factors

A general goal of plant hormone signalling is the activation or repression of transcription factors (TFs). TFs are proteins with special domains which bind DNA at specific regulatory sites usually found upstream of protein-coding regions, and activate/suppress the production of mRNA. The core TF binding sites (TFBS) are short, 5-8 base pairs long and variable: one or more nucleotide substitutions can be tolerated (Wray *et al.*, 2003). This means that the sequences for binding sites crop up at a high frequency at random within the genome due to their short length, which must make a significant proportion of these sites non-functional. Binding to the TFBS may take place in TF complexes (Brkljacic and Grotewold, 2017). This combinatorial gene regulation theory explains why TFs may bind to some cis-regulatory elements *in vivo* and different ones *in vitro*; this is due to indirect or cooperative binding through complexes (Hunt and Wasserman, 2014). Transcriptional regulation is the initial mechanism for regulating expression of genes and in turn affects the proteome, metabolome and phenome. TFs are master regulators of vital processes such as development, growth, metabolism and the response to the environment. TFs such as JAZs, MYC2, ORA59, ERF1, and EIN2 play an important role during infection by activating defence gene expression, suppressing processes such as photosynthesis and integrating responses from different signalling

pathways (Alonso *et al.*, 1999; Kazan and Manners, 2012, 2013; Nakano *et al.*, 2006; Pré *et al.*, 2008). Protein-protein interacting domains are frequently found in TFs - such as the helix-loop-helix motif in bHLH TFs, the leucine zipper in bZIP TFs and the ZIM domain in the TIFY family.

The sequencing of the Arabidopsis genome led to the identification of 1,500 genes in Arabidopsis encoding transcription factors (Riechmann et al., 2000) - over 5% of its genome. The number has increased in recent years as newer computational methods are used and currently stands at ∼2,000 TFs (Iida *et al.*, 2005; Pérez-Rodríguez *et al.*, 2009). Using newer annotated versions of the Arabidopsis genome and the known sequence of DNA-binding domains, Pérez-Rodríguez *et al.* (2009) estimated the number transcription factors to be ∼2,450 of ∼30,700 Arabidopsis genes - now around 8% of the genome - distributed amongst 82 TF families. Half of the TFs in Arabidopsis are plant-specific, with DNA-binding domains unique to plants; these include the AP2/ERF, NAC, YABBY, WRKY, GARP, TCP, SBP, ABI3-VP1, EIL and LFY families (Mitsuda and Ohme-Takagi, 2009). This abundance of unique TFs is probably a result of the sessile nature of plants - there is more pressure to adapt to the environment and these complex regulatory structures help with maintaining homeostasis. TFs (along with genes involved in signal transduction) are preferentially duplicated and retained in the Arabidopsis genome compared to other kinds of genes, highlighting their importance in the adaptability to stress and changing environments (Blanc and Wolfe, 2004).

TF superfamilies associated with the biotic stress response include MYBs (Dubos *et al.*, 2010), AP2/ERFs (Nakano *et al.*, 2006), WRKYs (Eulgem and Somssich, 2007), NACs (Puranik *et al.*, 2012), and bZIPs (Alves *et al.*, 2013). For instance, MYB30 is a positive regulator of the HR response (Vailleau *et al.*, 2002), MYB96 mediates crosstalk between ABA and SA signalling and leads to the synthesis of SA during infection (Seo and Park, 2010), MYB108 restricts the spread of necrotrophic pathogens by probably interacting with the jasmonate signalling pathway and ROS (Mengiste *et al.*, 2003), and MYB72 is involved in systemic resistance triggered by beneficial pathogens by mediating crosstalk between ET and JA signalling (Segarra *et al.*, 2009). ORA59, an AP2/ERF TF, activates the *PDF1.2* defence gene and integrates JA and ET signalling (Zarei *et al.*, 2011). ERF5 and ERF6 are involved in defence against *B. cinerea* (Moffat *et al.*, 2012). ERF2 and ERF4 act as activator and repressor, respectively, of the JA response in the defence against *Fusarium oxysporum* (McGrath *et al.*, 2005). ET and JA pathway converge on ERF1, causing its activation and action on further downstream defence response genes (Lorenzo *et al.*, 2003). The W-box motif bound by WRKY TFs is highly conserved in upstream regions of R genes and basal defence genes (Eulgem, 2005).

While some members of these TF superfamilies have been characterised, the role of most in biotic stresses is not known. Functional redundancy among family members means that typical knockout studies may not be able to deduce the importance of a TF in defence (Moore and Purugganan, 2005).

1.3.4.1 Plant gene regulatory networks

Gene regulatory networks (GRNs) capture the interactions between TF proteins and the downstream genes they regulate via binding to their promoter regions. Ferrier *et al.* (2011) reviewed analyses of genome-wide binding of specific TFs and found that TFs, on average, bind to 1200 target gene promoters. They estimated that, based on 27400 protein-coding regions and 1700 TFs, this would mean each Arabidopsis gene is regulated by an average of ∼75 different transcription factors. However, analyses focused on identifying the regulators of a single gene usually find a smaller number of binding events (Ferrier *et al.*, 2011). This implies that GRNs are also very dynamic: only a fraction of all potential interactions are active at any one time. The interactions active at a particular time differ based on environmental conditions and the plant tissue in question. This makes the elucidated network specific to a particular environment, developmental stage and time.

As expected, TFs can also regulate the mRNA production of other TFs, so complex regulatory networks arise between the TFs. This can also result in TFs indirectly influencing the expression of genes by regulating other TFs. There has been much emphasis on identifying the structure of these GRNs in order to predict how to manipulate them for a particular outcome such as stress tolerance (Hong-bo *et al.*, 2006), stress signalling (Singh *et al.*, 2002) and metabolite production (Iwase *et al.*, 2009). If the gene being engineered is a TF, it can result in the differential expression of all or some of the target genes it regulates, giving a wider reach than is possible by modifying non-transcription factor genes (Bhatnagar et al., 2007). This is in fact a similar strategy as that used by pathogens: both fungi and bacteria can secrete effectors that manipulate transcription factors such as JAZs and ERFs in order to downregulate the PTI response and undermine plant defence (Lo Presti *et al.*, 2015). In addition, the importance and scale of transcriptional reprogramming in the plant response to stress (Lewis *et al.*, 2015; Windram *et al.*, 2012) highlights the importance of transcriptional regulation. Elucidating these networks and further modifying them can help improve the plant's resilience to stress (Muhammad *et al.*, 2017).

Not all mRNAs are translated into proteins, therefore the relationship between mRNA and protein is not always linear. This is due to a number of other biological processes which can also affect gene regulatory networks in addition to transcriptional events. Following transcription, RNAs first need to be processed, which includes capping, splicing, cleaving and polyadenylation, and are then transported to the cytoplasm. All the RNA processing steps have bearing on downstream events such as transport, mRNA stability and translation (Lorković, 2009). Perhaps most importantly, up to two thirds of Arabidopsis genes can be differentially spliced (Ner-Gaon *et al.*, 2007), and the major isoform can change when plants are exposed to stresses (Staiger and Brown, 2013). Defective transport of RNA from the nucleus to the cytoplasm enhances the susceptibility to cold stress (Dong *et al.*, 2006).

mRNAs form complexes with proteins in the cytoplasm which are referred to as messenger ribonucleoprotein particles. In addition to being translated, these messenger ribonucleoproteins can be redirected to degradation pathways in P-bodies or temporarily

stored in stress granules. Plant stress can result in the additional sequestering of messenger ribonucleoproteins to P-bodies and stress granules (Weber *et al.*, 2008). Proteins can also function as chaperones, assisting with achieving the correct RNA conformation for splicing, transport and translation (Kang *et al.*, 2013). The levels of chaperones is affected by stress, and they also play an important role in growth and development (Yang and Karlson, 2011).

Short interfering RNAs (siRNAs) are important in regulating gene expression in plants in a process known as posttranscriptional gene silencing (Hamilton and Baulcombe, 1999) and are cleaved from double-stranded RNAs. For instance, the Arabidopsis *P5CDH* and *SRO5* genes form siRNA products which then downregulate the expression of *P5CDH* through its cleaving, enhancing ROS production and the salt stress response. This feedback loop is triggered by salt stress which promotes the transcription of *SRO5* (Borsani *et al.*, 2005). microRNAs (miRNAs) are similarly responsible for cleavage of transcripts or translational inhibition (Bartel, 2004), and differ from siRNAs by being generated from long hairpin precursors.

Post-translational modifications, such as phosphorylation, ubiquitination and sumoylation also affect the levels of protein, and therefore also have an effect on the dynamics of gene regulatory networks. These modifications affect a protein's location, function, activity, stability and therefore levels (Mazzucotelli *et al.*, 2008). However, in the construction of the networks in this work, post-transcriptional and post-translational processes which may affect the linearity of the mRNA-protein relationship are not taken into account due to lack of data and for preserving the simplicity of the models.

1.3.5 Transcriptional reprogramming

During the Arabidopsis stress response, many gene expression levels are activated or repressed in what is known as transcriptional reprogramming. Transcription factors, such as WRKYs, ERFs and bZIPs, are a large driving force of this phenomenon which is observed both during biotic (Windram *et al.*, 2012) and abiotic stresses (Kilian *et al.*, 2007).

Addition of phytohormones and MAMPs is enough to achieve this response, albeit at a lower level. Flg22 or oligogalacturonides treatment of Arabidopsis seedlings also causes a perturbation in 4413 genes 3 hours post treatment and 1672 genes 1 hour post treatment, respectively, with a return to normal levels after 24 and 6 hours, respectively (Denoux *et al.*, 2008). Prior to hormone signalling, defence genes such as *WRKY33*, *ERF1* and *TDR1* are rapidly upregulated in both treatments; this is followed by SA, JA and ET signalling, and a late response which consists of SA-regulated processes. Methyl JA treatment results in a wave of reprogramming of 4% of genes that represses the cell cycle progression and activates phenylpropanoid biosynthesis (Pauwels *et al.*, 2008). Approximately 22% of 23,000 genes were differentially expressed when Arabidopsis seedlings were exposed to chito-octamers or chitin (Ramonell *et al.*, 2005).

Identifying genes that undergo transcriptional reprogramming helps identify key players in the response. However, it is possible for TF levels to remain unchanged during a stress response but for them to play a key role - this can happen if post-

translational regulation like nuclear import signals are triggered by signals and change the TF equilibrium between the nucleus and cytoplasm. This has been observed with EDS1, a lipase-like protein that functions in R-mediated resistance, which is shuttled to the nucleus upon sensing of biotrophic and hemi-biotrophic pathogens where it mediates transcriptional reprogramming (García *et al.*, 2010) and AtbZIP10, which is shuttled to the nucleus upon infection to induce the HR and ROS-induced cell death (Kaminaka *et al.*, 2006).

1.3.6 *Botrytis cinerea*

One such pathogen that triggers the plant defence response is the fungus *Botrytis cinerea*. *Botrytis cinerea* is responsible for the common grey mould, and is an airborne necrotroph with a broad host range, infecting over 200 crop hosts worldwide, including Arabidopsis (Williamson *et al.*, 2007). It is capable of infecting different plant organs, from food storage organs (carrots and potatoes), leaves (lettuce, broccoli, cabbage), and fruit (strawberry, grape, raspberry, tomato), to flowers (rose, gerbera). As a result, the fungus was ranked the second most scientifically/economically important in a list composed by plant pathologists (Dean *et al.*, 2012). *B. cinerea* enters the plant at an early stage and can lay dormant until conditions are favourable and it senses that the host is mature or senescing. The outcome of this is apparently healthy crop which begins showing signs of damage after harvest, during transport or sale (Williamson *et al.*, 2007). Due to its ability to infect many different hosts, it is difficult to estimate the cost of *B. cinerea* damage; an estimated $66 million is lost a year in wine and table grapes in South Africa, Chile and Australia alone (Dean *et al.*, 2012). Such figures may well be underestimates as they do not take into account losses from dissatisfied customers who refrain from making future purchases due to previous bad experiences, or produce that was spoiled during transport.

B. cinerea is a robust pathogen, able to survive in diverse hosts, and in different states, such as mycelia, conidia or sclerotia in crop debris. Its genetic plasticity also contributes to the lack of effective control via fungicides (Leroux *et al.*, 2002). Defence against *B. cinerea* has no known R genes with a qualitative (all or nothing) effect and so involves many genes with quantitative effect (Denby *et al.*, 2004). However, there has been some success in identifying QTLs for resistance and introducing them from wild *Solanum* species into the cultivated tomato, *Solanum lycopersicum*, to give partial resistance to *B. cinerea*, reducing disease incidence by 48% (Finkers *et al.*, 2007).

1.3.6.1 *B. cinerea* infection process

Fungi use plants as a carbon source and to manufacture essential nutrients by releasing enzymes to digest macromolecules and absorbing the resulting nutrients through their hyphae. Necrotrophs cause vast mechanical and biochemical damage as they release toxic molecules and lytic enzymes to decompose plant tissue, unlike biotrophs which parasitise living tissue (Prins *et al.*, 2000). The main symptoms associated with *B. cinerea* are localised infection and expanding necrosis. The infection is usually initi-

ated by airborne spores, also known as conidia. In order to penetrate the host surface, *B. cinerea* conidia form structures called appressoria to breach the leaf cuticle. Conjecture that these appressoria secrete lipases and cutinases to aid in penetration was lessened by deletion studies of a lipase and cutinase that did not affect pathogen virulence (REIS *et al.*, 2005; Van Kan *et al.*, 1997) but there at least 17 more of these type of enzyme genes that could perform this function. The fungus facilitates host necrosis with phytotoxins such as botrydial and botcinolides but also by promoting the oxidative burst and the HR, thus using the plant's own defence system against it (van Kan, 2006). Oxalic acid is one such secreted molecule responsible for promoting programmed cell death by increasing ROS levels and inhibiting defence signalling (Williams *et al.*, 2011). Additionally, *B. cinerea* was found to hijack the Arabidopsis RNA interference machinery by secreting small RNAs that bind to the plant's Argonaute protein (Weiberg *et al.*, 2013). Argonaute, as part of the RNA-induced silencing complex, silences genes with sequences complementary to the fungal small RNAs at the transcriptional or post-transcriptional level. The targets of these small fungal RNAs include genes thought to play a role in host immunity: MAPKs MPK1 and MPK2, MAPKKK4, peroxiredoxin - a peroxidase potentially involved in ROS signalling - and a cell wall-associated kinase, WAK (Weiberg *et al.*, 2013).

1.3.6.2 The Arabidopsis defence response to *B. cinerea*

The plant response to *B. cinerea* is different to its response to biotrophic pathogens, as it needs to counteract the secreted phytotoxins and lytic enzymes that lead to apoptosis. Pectin is a major source of carbon for *B. cinerea*; in digesting it with endopolygalacturonase enzymes, oligogalaturonides (OG) are released (van Kan, 2006). OGs act as DAMPs and trigger the immune response as shown for MAMPs in Figure 1.2. Fungal endopolygalacturonases and chitin from the fungal cell wall act as MAMPs and trigger calcium influx, ROS production, MAPKs and phytoalexin production (Poinssot *et al.*, 2003). Arabidopsis membrane-bound receptor kinases CERK1 (Miya *et al.*, 2007) and LYM2 (Faulkner *et al.*, 2013) and the cytoplasmic receptor kinase BIK1 (Zhang *et al.*, 2010) recognise chitin while the receptor WAK1 (Brutus *et al.*, 2010) binds to OGs. These receptors associate with BAK1, and induce a MAPK signalling cascade that activates the WRKY33 defence gene which upregulates camalexin biosynthesis genes (AbuQamar *et al.*, 2017).

As with other biotic stresses, the phytohormones play a large role in the response to *B. cinerea* infection. JA signalling is vital for inducing the defence response to the necrotroph, as the increased susceptibility to disease of *jar1* and *coi1* mutants shows. Meanwhile, an increase in endogenous JA caused by the *fou2* mutation increases resistance (Bonaventure *et al.*, 2007). As might be expected from its aforementioned role in wounding, MYC2 knock-outs are less susceptible to *Botrytis* infection (Zhu *et al.*, 2011). ET is similarly important in defence, and its biosynthesis is triggered by pathogen detection. Plants lacking EIN2, EIN3 and EIL1 have higher *B. cinerea* infection ratios (Zhu *et al.*, 2011), while those constitutively expressing ERF1 show resistance to several necrotrophs (Berrocal-Lobo *et al.*, 2002). The role of SA signalling, meanwhile, remains

unclear. SA is required by the PTI to *B. cinerea* when it is triggered by flg22 (Laluk *et al.*, 2011); however, Arabidopsis mutants incapable of SA perception or synthesis do not show a modified response to infection (Ferrari *et al.*, 2003).

Transcriptional reprogramming has also been observed during infection of Arabidopsis leaves with *Botrytis cinerea*: 9,838 genes are differentially expressed as a result - about a third of the genome (Windram *et al.*, 2012). Genes that are responsible for ET and camalexin synthesis, ABA degradation, cell wall biogenesis and transport are among the genes upregulated, while genes connected with photosynthesis, and glucosinolate, chlorophyll and flavonoid biosynthesis, are downregulated.

1.4 The response to multiple stresses

The abiotic stress response follows a similar sequence of events as the response to biotic stresses. Stress perception is followed by signal transduction to the nucleus and cytoplasm, triggering a change in gene expression affecting the metabolome, and leading to acclimation. Transcriptional reprogramming also been observed in plants under salt, cold and osmotic stress, with approximately 2100, 1000 or 1100 genes of 8,100 genes studied being differentially expressed in cold stress, osmotic stress and salt stress respectively compared to the fresh medium control (Kreps *et al.*, 2002). Plants do not respond using insulated linear pathways for each kind of stress - these signals overlap, forming an intricate network that is responsible for the plant acclimation to the changes in its environment. Biotic and abiotic stress signalling pathways have many points of overlap, and can act antagonistically (Anderson *et al.* (2004), Yasuda *et al.* (2008), and see Figure 1.3).

In light of this, it is unsurprising that the response to combined synergistic stresses is more than the sum of the response to each individual stress (Rizhsky *et al.*, 2004). Predicting the plant's response to a combination of antagonistic stresses, such as cold and osmotic stress; or a mixture of biotic and abiotic stresses, which activate the antagonistic hormones salicylic acid and abscisic acid, respectively, is not an intuitive task (Sewelam *et al.*, 2014). To make up this deficit in knowledge, an increasing number of studies focus on plants that have been treated with two or more different stresses, including the interaction between biotic and abiotic factors (Cheong *et al.* (2002), Fujita *et al.* (2006), Suzuki *et al.* (2014)).

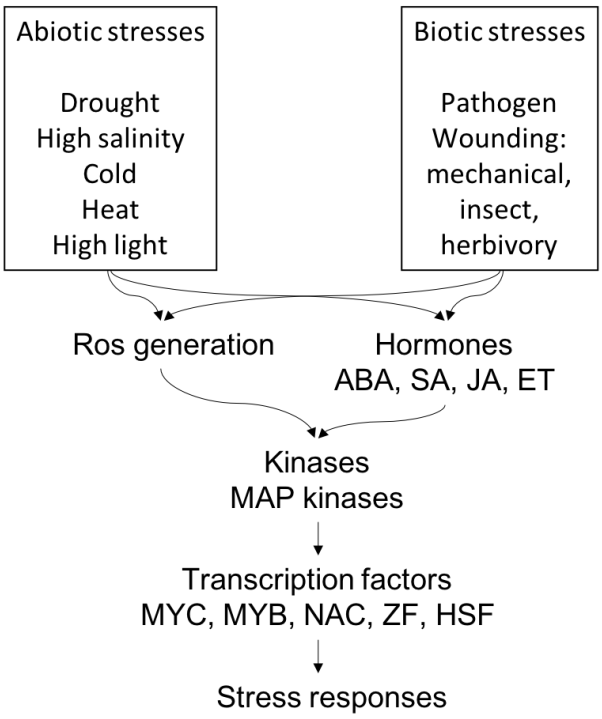


Figure 1.3: Points where the biotic and abiotic stress signalling networks converge. Crosstalk has been observed between ROS, phytohormones, signalling cascades and transcription factors triggered as a response to biotic and abiotic stresses, implying that plants integrate signals from multiple sources which allow them to respond to abiotic stresses and diseases. Figure adapted from Fujita *et al.* (2006).

1.5 Systems biology approaches for capturing the plant stress response

The large-scale behaviour of particles can be counter-intuitive and surprising, but not necessarily unpredictable, given the use of appropriate quantitative models. As discussed above, a complex response like defence in plants is not limited to changes in a single or a few proteins, mRNA transcripts or metabolites. Rather, it is a coordinated, systemic response involving changes in gene expression and regulation, signalling pathways and biological networks, and requires a comprehensive and high-throughput study approach to fully appreciate the impact of each molecular player. A quantitative systems or 'omics' approach to understanding normal function and the disease response involves large scale characterisations of gene expression (transcriptomics), protein levels (proteomics), metabolic levels (metabolomics), epigenetic modification (epigenomics), genome variants (genomics) or the full set of physical and biochemical traits of an individual (phenomics). These studies favour an interdisciplinary approach - both biological and computational/modelling tools are needed for gathering and analysing data and

building models to understand the system.

Systems-level analyses can be applied to fully understand the workings of a single cell (Libault *et al.*, 2017) or conducted at the tissue, organ or entire plant level (Lucas *et al.*, 2011). At the subcellular scale, studies are divided into the bottom-up or top-down approach (Bruggeman and Westerhoff, 2007). The top-down approach collects omics data and uses it to identify molecules or the underlying interactions and mechanisms. Bottom-up approaches integrate highly detailed data of constitutive parts to form a systems-level description of a process. Subcellular systems biology has previously been used to study root hair cell differentiation, development and function (Libault *et al.*, 2010), and signalling events in guard cells triggered by environmental changes (Raghavendra and Murata, 2017). At the tissue scale, cells are considered nodes and their physical/mechanical interactions with other cells as well as their exchanges of molecules, such as auxin, are modelled. Root patterning (Mironova *et al.*, 2010), auxin transport (Twycross *et al.*, 2010) and veination (Bayer *et al.*, 2009) have all been modelled at the tissue level. Leaf senescence has been studied at the organ level over 23 days by measuring leaf chlorophyll and total protein levels, and daily microarray profiling followed by clustering of gene expression to reveal functional gene clusters and identify groups of coregulated genes (Breeze *et al.*, 2011). Whole plants can be described using ecophysiological models that consider the plant as an organism presenting a phenotype, which is being influenced by and interacting with its environment. For example, the temperature response of developing maize, rice, and Arabidopsis plants was studied by looking at the rate of developmental processes and photosynthesis (Parent *et al.*, 2010). The structural development of plants can also be mathematically modelled using architectural models such as Lindenmayer systems (Prusinkiewicz and Lindenmayer, 2012) and multiple tree graphs (Guédon *et al.*, 2001).

As well as providing a unified view of biological processes, systems approaches can also be used to make predictions about the properties of a system and its components, infer interactions and also for optimisation purposes - and hence for synthetic biology. For instance, Zhu *et al.* (2007) have predicted how to increase C_3 photosynthesis by 76% by changing the distribution of resources among the enzymes of photosynthesis rather than the total amount of resources. Models have also been used to suggest improvements to plant breeding. Chenu *et al.* (2009) took a mixed modelling approach to explain the effect of genotype and the environment on growth phenotypes in maize. Their simulations took into account genetic inputs resulting in specific traits and environmental inputs such as soil conditions, daily weather and human inputs such as irrigation and sowing dates. This approach was then used to successfully predict the effects of QTLs on leaf and silk elongation in maize in conditions of drought.

1.6 Systems biology approaches for modelling gene regulatory networks

A predominant use of omics data is for constructing gene regulatory networks (GRNs) to understand, predict or influence the relationship between TFs and their

targets. This relationship is represented in network models by nodes (genes) and edges for the relationship between nodes. The edges in GRNs represent either positive or negative regulation in directed networks, or unspecified regulation or co-regulation in undirected networks. Undirected edges imply coexpression while directed edges represent a causal (direct or indirect) regulatory relationship, with the initial node being the regulator. Indirect relationships include the regulator influencing the target through other unknown intermediaries. In directed networks, edges imply causality.

One application of directed networks is as a framework for modelling the system with ordinary differential equations (ODEs). ODEs model the relationship between components such as mRNA, proteins and complexes, which are present in a large number. Due to the detail captured by ODEs, they require accurate information about the reaction, formation and degradation rates of the molecular species. ODE models have been used to describe Arabidopsis floral organ formation (van Mourik *et al.*, 2010), vascular patterning in roots (Muraro *et al.*, 2014) and carbohydrate metabolism (Nägele *et al.*, 2010) to name a few. The roots of mathematical biology, as it this is known, go deep into the 18th century, where ODEs were used to model population growth by Thomas R. Malthus.

A well established GRN that has been used to build a refined model from experimental data is the Arabidopsis circadian clock (Locke *et al.*, 2006; Pokhilko *et al.*, 2010). Pokhilko *et al.* (2010) extended the previous model for a final ODE system consisting of 28 ODEs, with parameters known from previous experiments and inferred from time series data. The model informed a new hypothesis that was then experimentally validated: a function for PRR5 as a night inhibitor for *LHY* and *CCA1*, and produced simulations consistent with biological data for entrainment and Arabidopsis *toc1*, *ztl* and *lhy/cca1/gi* knockouts.

One advantage of using systems of ODEs - coupled differential equations that arise when modelling GRNs, where several genes in the system can influence the levels of several others - is that there are many readily available algorithms to solve the differential equations and others to analyse the properties of the system, such as stability. Linear ODEs, in particular, are simple to state and simple (but sometimes time-consuming) to find solutions for. When modelling certain effects such as genetic toggle switches, discrete stochastic methods are preferred as they capture the dynamic behaviour of the system, which ODEs are unable to (Tian and Burrage, 2006). As deterministic models, such as ODEs, can only describe the average behaviour of a population, they cannot capture the cases when noise in the bistable system causes state switching. Stochastic methods deal with numbers of particles (discrete) rather than concentrations (continuous), and as the name implies, are governed by stochastic laws, making them well-suited for capturing the effects of noise in molecules expressed at low levels. An additional advantage that ODE models have over stochastic models is down to the computational effort needed to analyse stochastic models. It is also possible to approximate stochastic behaviour using a deterministic system in some cases; this has been shown using continuous and stochastic Petri nets (Heiner *et al.*, 2008), and stochastic differential equations, which have a propensity function that generates the time and index of the next occurring

reaction (Tian *et al.*, 2007). Therefore, an understanding of the system one wished to model is key in determining the best mathematical formulation of the model.

1.6.1 Data for inferring GRNs

GRN models can be built from biological data that provides direct evidence for a TF-target interaction. A simple way to determine this *in vivo* is to perform a yeast-one-hybrid (Y1H) experiment. Y1H is a bottom-up approach used for screening a collection of TFs against a particular promoter to find potential regulators. As the name implies, the work is carried out in *S. cerevisiae*, usually with a chromosomal *his3* mutation. The TF is fused to a GAL4 (a yeast transcription activator) domain and the DNA sequence of interest, or bait sequence, is placed before the coding region of the *HIS3* growth marker. If the TF binds to the bait, GAL4 is able to activate *HIS3* production, complementing the mutation and allowing the yeast to grow. Lack of an interaction between the TF and the bait results in no yeast growth (Ouwerkerk and Meijer, 2011).

However, due to the TF being artificially expressed in yeast, spurious interactions may occur and there is no guarantee that this binding occurs in the native cell under a particular condition such as stress, a particular developmental stage or a particular organ/tissue. An improvement upon yeast-one-hybrid is the use of ChIP-CHIP (Chromatin immunoprecipitation followed by hybridisation to microarray chips) or ChIP-SEQ (Chromatin immunoprecipitation followed by sequencing) to directly resolve the DNA sites that transcription factors bind to in a promoter at a particular time *in planta* (Kaufmann *et al.*, 2010). ChIP isolates a DNA-binding protein of interest (such as a transcription factor, DNA methylase or histone), together with the DNA it is bound to. This is done by cross-linking the protein to DNA and purifying it using an antibody specific to that protein. ChIP-SEQ and ChIP-CHIP are therefore carried out for one or a few DNA-binding proteins at a time, thereby providing detailed information on a single or a few TFs. Owing to the nature of the method, and the low availability of antibodies for proteins, doing so for thousands of transcription factors simultaneously under different stress conditions presents great difficulty technically and financially. Even so, TFs have also been found to bind to hundreds of different DNA fragments, only a few of which respond transcriptionally to the TF (Hornitschek *et al.*, 2012; Yant *et al.*, 2010). Unsurprisingly, it was determined that TFs can bind to TFBS which are completely different from the primary TFBS (Franco-Zorrilla *et al.*, 2014).

Experimentally, more high throughput approaches involve protein binding arrays (Gong *et al.* (2008), Godoy *et al.* (2011)) and DNA-affinity purification sequencing (O'Malley *et al.*, 2016). Computationally, TFBS can be determined by compiling sequences of promoters coregulated by a particular TF and looking for an overrepresented DNA motif within them. There are a number of algorithms that do this such as MEME (Bailey *et al.*, 2015), Pscan (Zambelli *et al.*, 2009), and MatInspector (Cartharius *et al.*, 2005). As illustrated by sequence logos, binding is not strictly limited to a specific sequence, however, there are certain nucleotides in the TFBS that are strongly conserved, where the TF appears to not tolerate variation. Computational methods lead to a large number of false positives in the form of non-functional sites - sites that have the TFBS

but do not impact gene regulation. This is due to the large amount of noise in the system: binding sites are smaller than 10 nucleotides long and promoter regions can span 1-2 kilobases (kb) or more. In addition, many members of plant TF families show preferences for the same DNA motif, while maintaining their individual functions - for instance most WRKYs bind to the W box domain (T)(T)TGAC(C/T) (Eulgem *et al.*, 2000) while the core sequence recognised by the bHLH TFs is CANNTG (Toledo-Ortiz *et al.*, 2003), but each TF can have very different response once bound, and not all TFs from that family regulate each binding site.

Another method of inferring regulatory links is with the help of knock-out (KO) and overexpression (OE) data. KO data is obtained by creating a homozygote plant with a particular gene disabled through mutagenesis. This allows the study of subnetworks directly and indirectly affected by that gene. For instance, if a TF is an activator or a repressor of genes downstream of it, then the downstream gene levels should be decreased/increased in a knockout in comparison to the wild type expression. This is the simplest case; it is possible for redundancy to mask the effect of a KO on the genes downstream of the regulator. This has been observed, for instance, in the circadian regulators, LHY and CCA1 (Mizoguchi *et al.*, 2002). Perturbations of a system by transiently increasing or decreasing the amount of regulators can provide additional information that can help uncover redundant connections without the drawbacks of inactivating a gene for the duration of the plant's life.

1.6.1.1 GRNs derived from transcriptomic data

Systems-level monitoring of organisms is made possible by improvements in high-throughput methods. However, quantifying the global amount of TF proteins available in a cell or tissue at a particular time is still economically infeasible and time-consuming, so the amount of mRNA transcripts in the cell is often used as a proxy for the protein concentration. Increasingly, complex processes are being deciphered through the generation of high-resolution temporal transcriptomic data, as they can provide a snapshot of the current state of the transcriptome in a tissue (Breeze *et al.*, 2011; Covington *et al.*, 2008; Loraine *et al.*, 2013; Marquez *et al.*, 2012; Sparks and Benfey, 2017). mRNA quantification can be done using microarrays, RNA-seq and NanoString for hundreds of genes at the time with automated systems and minimal input from humans. Transcriptomic data provide a wealth of information that can be used by algorithms to predict the regulatory links between different genes, as described in more detail below.

The large quantities of data also make inferring networks something that necessitates computational power in the form of network inference algorithms. The amount of a particular mRNA depends on the activity of its regulators, the leaky transcription rate, the decay of the mRNA, randomness, and measurement errors (as described in Section 1.3.4.1) - some or all of these variables can be considered by the network inference method. Due to the multiple influences from sources other than TFs, integrating time series data with other types of static data such as knockout and perturbation data, protein and metabolome data can be very informative (Hecker *et al.*, 2009; Yip *et al.*, 2010). Studies in model prokaryotic and eukaryotic systems have shown that the inte-

gration of such priors based on data other than gene expression has resulted in more accurate networks than the sole use of gene expression data (Arrieta-Ortiz *et al.*, 2015; Greenfield *et al.*, 2013).

The simplest way of constructing a network using transcriptomic data is by clustering genes that show a similar expression pattern with the assumption that coexpression implies coregulation; this is known as the "guilt-by-association" heuristic (see Wolfe *et al.* (2005) for a review of guilt-by-association applicability). Coexpressed genes also show higher than expected functional coherence (Wu *et al.* (2002), Stuart *et al.* (2003)). Simple correlation, like Pearson or Spearman, or other correlation approaches like Gaussian graphical models (Schäfer and Strimmer, 2004), measures the linear dependency of genes, so that non-linear dependencies will not be spotted. A more sophisticated kind of coexpression network is based on information theoretic approaches like mutual information, derived from gene expression profiles (Butte and Kohane, 1999). Mutual information is a measure for how much one variable is dependant on another. Mutual information-based algorithms, like ARACNE (Margolin *et al.*, 2006), CLR (Faith *et al.*, 2007) and MRNET (Roth *et al.*, 2003) build fully connected graphs which then require thresholding to filter out weak edges and indirect interactions.

The networks inferred by guilt-by-association are generally symmetric - the techniques capture bidirectional links between connected nodes (Markowetz and Spang, 2007). As a result, lots of regulatory mechanisms can explain coexpression (see Figure 1.4) and biologically relevant motifs such as feedback and feedforward loops cannot be resolved. Nevertheless, coexpression networks have been used to predict known root transcriptional modules and identify new TFs in the regulation of secondary cell wall synthesis in the root (Montes *et al.*, 2014). They are also useful for predicting function of uncharacterised genes in agriculturally important species (Schaefer *et al.*, 2017). While coexpression networks can be used to form hypotheses generation and inferring the function of genes through guilt-by-association, they are not capable of predicting gene levels in novel conditions, and therefore were not used in this analysis. Furthermore, simple correlation and mutual information methods did not perform very well at inferring gold standard networks, compared to other methods employing regression, ANOVA, decision trees or multiple approaches in a large survey of network inference algorithms (Marbach *et al.*, 2012).

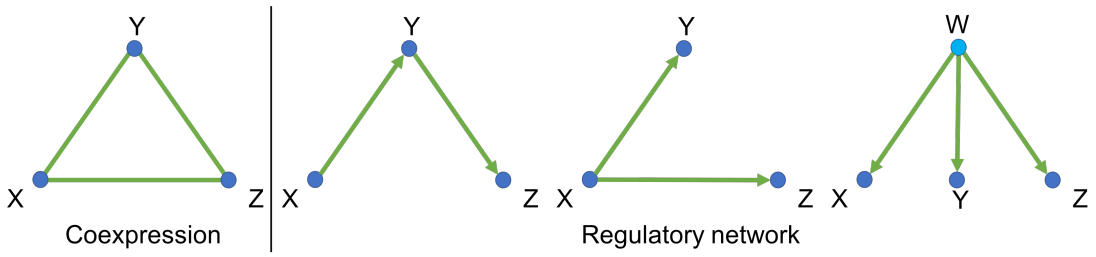


Figure 1.4: Different mechanisms can explain coexpression, with some of the simplest shown here. The leftmost plot in the dashed box shows three coexpressed genes in a module from a coexpression graph. Possible regulatory relationships that can explain the coexpression are shown in the other three plots: gene regulation takes the form of a cascade (left), or one gene regulates the two other genes (middle), or there is a common 'hidden' regulator (right), which is not part of the model. Figure adapted from Markowetz and Spang (2007).

A second approach to solving network structures is using Boolean networks. This is the simplest method for recording gene states in GRNs, which are represented using Boolean logic (0 for off/absent or 1 for on/present) at a particular time point (Shmulevich *et al.*, 2002). Logical operators such as "AND", "OR", and "NOT" are used to describe the interactions between different genes, as shown in Figure 1.5. Boolean networks were used by Saint Savage *et al.* (2008) to uncover the GRN responsible for patterning in Arabidopsis root epidermis and extend the previous network models to include TFs CAPRICE and GLABRA3 as the orchestrators of this patterning. Their models and subsequent experiments also showed a limited role of the TF WEREWOLF compared to prior conceptions. Boolean networks require the discretisation of the gene expression data, which results in information loss. For instance, the predominant type of data used in this work is obtained from microarrays, which, when normalised, takes on values between 5 and 12. The first step in converting this to a Boolean network is to decide an arbitrary threshold for the cut-off between on and off states. Furthermore, because we are interested in the change in expression of a gene over time as a result of stress, the categorisation of a gene requires at least 3 states: 0 = 'unchanged', 1 = 'increase', and -1 = 'decrease'. Ultimately, for predicting the effects of regulatory rewiring on genes in a GRN, the amount of change in gene expression is an important factor, rather than just the on/off state in a gene, which is why Boolean networks were not chosen for this work.

Gene X	Gene Y	Gene A
0	0	0
0	1	0
1	0	0
1	1	1

Gene X	Gene Y	Gene A
0	0	0
0	1	1
1	0	1
1	1	1

Gene X	Gene Y	Gene A
0	0	1
0	1	0
1	0	0
1	1	0

Gene X	Gene Y	Gene A
0	0	0
0	1	1
1	0	0
1	1	0

Figure 1.5: Truth tables for possible regulatory interactions in a Boolean network model. The target gene is Gene A, and its regulators are Gene X and Gene Y. Top left: "AND". Gene A is only switched on when both Gene X and Y are switched on. Both X and Y are positive regulators of A. Top right: "OR". Gene A is switched on when either Gene X or Gene Y are on. Both X and Y are positive regulators of A. Bottom left: "NAND". Gene A is only switched on when Genes X and Y are switched off. In this case, both X and Y are negative regulators of A. Bottom right: "Y, NOT X". Gene A is only switched on when Gene Y is on and Gene X is off. X is a negative regulator of A that takes precedence over the positive regulator Y.

Bayesian networks with direct acyclic graph structures have also been used to generate graphs for regulatory networks (Ben-Gal, 2008). They combine prior knowledge with gene expression data in a probabilistic framework. Modelling with Bayesian networks usually involves building a network that best explains the data (model selection) followed by estimating a conditional probability distribution at each node (parameter learning). The Aikake information criterion or Bayesian information criterion are used for model selection (Liu *et al.*, 2012). While static Bayesian networks are directed networks, they cannot include cycles (feedback) by definition. Dynamic Bayesian networks (Beal *et al.*, 2004) can be used to circumvent the limitation of Bayesian networks with regards to feedback loops, but they are still limited to small network sizes due to the problem of computing the likelihood for all possible network structures. Needham *et al.* (2009) derived a network of the circadian clock using a Bayesian network-based inference. They then used this network to propose that GATA TFs also functioned in the photosynthetic network.

Regression-based inference approaches use feature selection to identify the regulators that best explain the expression levels of a target gene. Different feature selection approaches can be used, such as least angle regression (Efron *et al.*, 2004), ridge regression (Hoerl and Kennard, 1970) and least absolute shrinkage and selection operator (Tibshirani, 1996). The former appears to be the most popular feature selection as it re-

sults in a sparse network with few regulators per node (Marbach *et al.*, 2012). Yao *et al.* (2011) constructed a regression model to provide a better understanding of the photosynthetic light acclimation in Arabidopsis from time series datasets. These regression methods are ultimately still linear models and therefore cannot handle combinatorial and non-linear interactions. Non-linear regression models based on decision trees (Huynh-Thu *et al.*, 2010), or Gaussian processes (Lawrence *et al.*, 2007; Penfold and Wild, 2011) have also been proposed to capture non-linear and combinatorial interactions.

A further review and ranking of network inference algorithms is provided in Marbach *et al.* (2012), and the strengths and weaknesses of different mathematical approaches are discussed in Schaffter *et al.* (2011).

1.6.1.2 Correlation of mRNA and protein levels

Quantifying the amount of TF proteins available in a cell at a particular time is difficult, so the amount of mRNA transcripts in the cell is often used as a proxy for the protein concentration, and for inferring networks as described above. This procedure is not without its disadvantages as protein levels do not always correspond to transcript levels; this method overlooks the processes of translation, protein degradation and post-translational modification. These methods provide a snapshot of the current state of the transcriptome in a tissue. However, metabolites are short lived and therefore hard to quantify. Correlation between mRNA transcript levels and protein levels has been previously been described as poor (Table 1.1).

Table 1.1: Pearson correlation coefficient quantifying the linear relationship between the mRNA and protein expression values.

Organism	Pearson correlation coefficient	Dataset size	Reference
S. cerevisae	0.66	2044	Greenbaum *et al.* (2003)
S. pombe	0.58	1367	Schmidt *et al.* (2007)
E. coli	0.57	1103	Ishihama *et al.* (2005)
M. musculus	0.59	425	Tian *et al.* (2004)
Arabidopsis thaliana	0.51	1176	(Lan *et al.*, 2012)
Zea mays	0.624	2813	(Ponnala *et al.*, 2014)

The path from mRNA to protein is complicated by other mechanisms such as RNA secondary structure, translational modulators, codon bias, ribosome occupancy on the mRNA. Experimental error and noise also play a big part in the observed discrepancy between mRNA and protein values (Maier *et al.*, 2009). However, this disparity is largely accounted for by differences in half life between mRNA and proteins in *S. cerevisae* (Wu *et al.*, 2008a). Protein turnover rates are in turn determined by post translational modification, their subcellular localisation, the protein's intrinsic stability and the N-end rule (Maier *et al.*, 2009).

A recent large-scale study in 29 human tissues spanning 12,894 genes with detectable transcript and protein levels has found that in 90% of cases, mRNA and protein levels do strongly correlate (Wang *et al.*, 2018). In this dataset, the dynamic range of

transcripts is 4 orders of magnitude and the dynamic range of proteins spans 8 orders of magnitude, making the relationship between mRNA and protein levels nearly quadratic in all tissues. However, the authors raised the point that a potential reason behind the discrepancy is the way mRNA and protein levels are quantified. With RNA quantifications such as RNA-Seq, reads that hit multiple transcripts are assigned to all those transcripts, while peptide fragments are usually assigned only to the protein with the most evidence. Mass spectrometry also typically underestimates the levels of proteins that are highly hydrophobic or have a large number of transmembrane domains, unless specialised methods are employed (Friso *et al.*, 2004). Incorrect annotation - genes annotated on the wrong strand or to inaccurate loci, and pseudogenes wrongly predicted as genes - can also play a part in the inaccurate prediction of protein levels (Peltier *et al.*, 2004; Ponnala *et al.*, 2014). Similarly, a study by Li *et al.* (2014) that rescaled proteome data from Schwanhäusser *et al.* (2011) found that transcription plays a more important role than translation in determining protein levels. Using the new protein abundance values, 84% of the variation in protein levels could be explained by mRNA levels, compared to only 40% in the original study.

Overall, mRNA levels are a good proxy for protein levels as recent studies are beginning to show - although this remains to be seen in cells not in their resting state. Xu *et al.* (2017) found that pattern-triggered immunity upon exposure to the MAMP elf18 leads to modification in translation efficiency which were not very well correlated with the levels of mRNA.

1.6.2 Rewiring of GRNs

Transcriptional activity has proven to be very important in plant domestication and diversification - 62% of the loci associated with domestication phenotypic changes belong to TFs (Meyer and Purugganan, 2013). Of these, 43% of the changes are rewirings in the regulatory regions affecting the expression of the TFs. Network rewiring is the rearrangements of connections in a network and can consist of edge loss, edge gain or swapping of edges. Modification of cis-regulatory elements of a gene can change its regulation and therefore its expression levels. By combining different promoter regions with the open reading frame of transcription factors, it is possible to control when the TF is expressed, and by how much. This will, in turn, control the expression levels of genes downstream of the transcription factor. This dynamic process has been observed during evolution, cell specialisation, and the response to stress, as detailed below (Ideker and Krogan, 2012).

1.6.2.1 Rewiring in evolution

The evolution of morphological traits is generally led by rewiring of connections in gene regulatory networks (GRNs) (Gompel *et al.*, 2005). Mutations in regulatory elements result in modification of gene expression, and these are better accommodated than mutating the actual coding sequence, particularly in highly pleiotropic genes (Prud'Homme *et al.*, 2006). Rewiring the regulation of a gene is a powerful tool

for rapid adaptation to different ecological niches, as evidenced by the fact that it can result in more divergence between related species than actual sequence variation in gene coding regions (Borneman *et al.*, 2007). Following genome duplication in *Saccharomyces cerevisiae*, a certain cis-regulatory motif was lost from a number of genes, leading to rewiring en-masse, which is hypothesised to have caused the ability of *S. cerevisiae* to grow anaerobically (Ihmels *et al.*, 2005).

On a larger scale, whole genome duplication (WGD) events followed by rewiring commonly develops genes with novel functions. Genome duplication provides a selective advantage to the organism as it contributes to innovative trains and an increase in biological complexity (De Smet and Van de Peer, 2012). At the same time, the network redundancy provides a safety net for functional innovation by increasing the robustness to mutations (Aldana *et al.*, 2007). Subfunctionalisation and neofunctionalisation driven by rewiring follow genome duplication. For instance, duplication of Arabidopsis CUP-SHAPED COTYLEDON (CUC) genes gave rise to the homologous genes CUC1 and CUC2 (Hasson *et al.*, 2011). CUC1 and CUC2 regulate different target genes and are in turn regulated by different genes and as a result, have different functionalities with respect to leaflet formation (Hasson *et al.*, 2011; Spinelli *et al.*, 2011). Rewiring can also change spatiotemporal patterns, such as the case of the homologs FLOWERING LOCUS T1 and T2 in the poplar which differ in their expression throughout the year (Hsu *et al.*, 2011) and *Brassica napus* ribosomal homologs have tissue specific expression patterns as a result of subfunctionalisation following a WGD event.

Evolution takes place over many generations; however, rewiring can also act on a single generation in an organism. Different environmental conditions show that interactions within a cell are not static and can be altered by stress such as DNA damage (Bandyopadhyay *et al.*, 2010).

1.6.2.2 Rewiring of TF GRNs

TF rewiring has been shown to be a powerful tool in the evolution of domestic crops (Century *et al.*, 2008; Doebley *et al.*, 2006). The rewiring of the TCP TF Tb1 produces the main difference between maize and its wild ancestor, teosinte. Through modifications of *Tb1*'s promoter region, the maize plant develops short branches with ears at their tips, rather than the long branches tipped by tassels like teosinte (Wang *et al.*, 1999). A key trait in rice has also arisen from a single nucleotide polymorphism in the regulatory region of the gene *qSH1*. This changes the expression pattern of *qSH1* in the abscission layer of the rice seed, preventing it from dropping off the panicles - known as grain shattering - and increasing harvesting efficiency (Lin *et al.*, 2007). Similarly, changes outside of the protein-coding region of TF *fw2.2* in wild tomato leads to lower levels of expression and enhanced fruit growth (Cong *et al.*, 2002). Other regulatory changes underlying varietal differences such as kernel colour in maize and carotenoid content in maize are summarised in Doebley *et al.* (2006).

Extreme changes in the regulation of transcription factors - knockouts and over-expressions - can have large effects on plant phenotypes (Jiang, 2004; Lai *et al.*, 2008; Umezawa *et al.*, 2006; Zheng *et al.*, 2006). However, while KOs and OE can enhance

the stress response, such extreme measures can cause significant stress if the knocked out gene is required for other processes or by the overexpressed gene being a drain on resources or interfering with other processes (Karim *et al.*, 2007). This usually leads to growth retardation (Abe *et al.*, 2003; Hsieh *et al.*, 2002; Kasuga *et al.*, 1999; Liu *et al.*, 1998). By contrast, using a stress-inducible promoter rather than a constitutively overexpressing promoter, the effects on plant growth are minimised (Kasuga *et al.*, 1999).

Rewiring can be employed as a tool to engineer organisms with optimised pathways and therefore outputs. Cellulose from plant cell walls is predominantly used in the paper industry and has other potential applications in biofuels and chemicals industry (Boisen *et al.*, 2009; Carere *et al.*, 2008). However, strategies for increasing cell wall thickness and reducing recalcitrance by lowering lignin content have mostly resulted in severe loss of biomass (Shadle *et al.*, 2007; Voelker *et al.*, 2010). Yang *et al.* (2013) used synthetic biology tools to rewire pathways from the secondary cell wall network to reduce lignin content and increase cell wall deposition. They introduced a positive feedback loop by adding a second copy of the *NST1* gene under the control of the *IRX8* promoter. They also removed the regulation of lignin biosynthesis *C4H* gene by NST1 by introducing a new copy of *CH4* regulated by the *pVND6* promoter in a *ch4* knockout. As a result, their plants had increased polysaccharide deposition but decreased lignin, making enzymatic hydrolysis more efficient.

Rewiring has emerged as a powerful and precise tool to achieve desired changes in a system, harnessed by evolution and biological engineers alike. Taking advantage of the large amount of data regarding the Arabidopsis response to *B. cinerea* attack, this work attempts to build a framework incorporating this knowledge which can be used to optimise Arabidopsis GRNs for defence against the necrotroph.

1.7 Project aims

- Elucidate GRNs underlying Arabidopsis response to stress

Synthetic biology greatly benefits from a quantitative and modelling approach for solving problems. The process of gene engineering in plants is slow, so narrowing down the number of potentially important TFs to modify using insights gained from network modelling is important. Therefore, the first step involves inferring accurate, experimentally-backed transcription factor networks from Arabidopsis plants in different stress conditions using network inference algorithms. The network structures are then closely examined for any emergent properties (Chapter 2).

- Design of improved defence network

The 'design' part of the design-build-test approach involves simulating the Arabidopsis-*B. cinerea* networks derived in the first part with two different systems - linear ordinary differential equations (Chapter 3) and Gaussian processes (Chapter 4). As rewiring has been proven to be a main driver of quick adaptation and phenotypic change, this work

will use rewiring in combination with these modelling techniques to find an optimised version of the plant network that enhances the defence response.

- Harnessing protoplast systems for building and testing rewirings

Finally, a framework for designing and carrying out rewirings in Arabidopsis plants and protoplasts is tested in Chapter 5. The aim of the rewiring is to enhance the defence response of Arabidopsis to *B. cinerea*. The merits of first testing rewirings in protoplasts as a high throughput system are explored, followed by a comparison of these results with stably rewired Arabidopsis plants.

Chapter 2

Elucidating Gene Regulatory Networks Underlying Arabidopsis Stress Response

2.1 Aims

This chapter aims to identify the best network modelling algorithms for inferring Arabidopsis transcription factor gene regulatory networks (TF GRNs) and furthermore use the obtained best network to explore the relationship between network structure and function.

- Infer individual TF networks using top network inference algorithms from data of Arabidopsis exposed to biotic or abiotic stresses.
- Rank the network inference algorithms based on their performance using *in silico* network data, for which the true network is known.
- Explore the potential of using the Bayesian Information Criterion as a way to rank network models.
- Rank the network inference algorithms based on their performance on a highly-validated *in vitro* network, for which the gold standard is constructed from experimental data.
- Rank the network models obtained from the time series of Arabidopsis infected with *Botrtyis cinerea* using experimental data which provides evidence for TF-gene binding *in vivo*.
- Determine the best network for use in Chapters 3 and 4.
- Use the best network model (or the consensus) to determine whether any node or network characteristics are linked to the increased likelihood of a TF being a positive or negative regulator of defence.

- Identify a core Arabidopsis stress network.

2.2 Introduction

The deduction of genome-level regulation has been made feasible by the widespread availability of transcriptomic (particularly time series) datasets. At the transcriptomic level, the challenge of reconstructing the regulatory links between transcription factors and their target genes is enormous, and needs to be solved with the aid of computational biology. Numerous network inference algorithms are available to construct networks out of gene expression data (Banf and Rhee, 2017; Marbach *et al.*, 2010, 2012). The performance of most of these is usually tested on small *in silico* datasets which are dissimilar to large-scale GRNs. This is largely because experimentally-validated gold standard networks for benchmarking are unavailable, other than for a few well-characterised small pathways such as the core network in human embryonic stem cells (Chen *et al.*, 2008; Kim *et al.*, 2008a) and the *E. coli* transcriptional regulatory network (Keseler *et al.*, 2010; Salgado *et al.*, 2012). In plants, GRN inference characterisation is ongoing, even for well-studied systems such as the Arabidopsis circadian clock (Fogelmark and Troein, 2014; McClung, 2006) (and see Bujdoso and Davis (2013) for an evolution of the clock models), and Arabidopsis flower formation (Ó'Maoiléidigh *et al.*, 2014; Pajoro *et al.*, 2014). At the same time, methods that perform well on a dataset from one organism or source, may perform poorly when inferring networks from a different dataset (Marbach *et al.*, 2012). Given this uncertainty, five different algorithms were used in this work, and the resulting models were ranked using various measures, before selecting a network model for further work.

Three of the five algorithms were chosen as they were top of their category in a study performed by Marbach *et al.* (2012) (GENIE3, Inferelator, TIGRESS). In addition, two other methods: CSI and GRENITS (Penfold and Wild, 2011) were also used for network inference due to their good performance on the same *in silico* datasets as tested in Marbach *et al.* (2012). The 5th DREAM network challenges were designed to identify the best network inference algorithms through a comprehensive characterisation. Network inference experts used their programs to make predictions on benchmark *in silico* and *in vivo* datasets. Marbach *et al.* (2012) identified six main categories of inference approaches; the accuracy of programs within these categories varies widely. The categories are: regression, mutual information, correlation, Bayesian networks, meta predictors and other predictors. A community network was also formed using the predictions of all participating experts. Network inference performance was ranked using the area under the precision-recall (AUPR) curve for all test datasets. TIGRESS had the best performance of all regression methods, Inferelator had the best performance of all meta predictors and GENIE3 was the best performer from the other predictor category. A short description of each algorithm follows, with further detail of how the problem of network inference is stated and solved mathematically given in Appendix A.

CSI takes a non-parametric approach to learning data for non-linear dynamic systems (Penfold and Wild, 2011). In the dynamic gene regulatory model, the com-

bined expression of upstream genes, called parents, directly or indirectly regulates the expression of downstream genes. The method uses Gaussian process dynamical systems (GPDS) to identify the relationship between gene profiles, capturing the network dynamics in a way that enables quantitative predictions. Gaussian process dynamical systems are further expanded on in Chapter 4 of this thesis.

GENIE3 is a network inference tool based on Random Forest or Extra-Trees methods for solving regression problems (Huynh-Thu *et al.*, 2010). Random Forest (the method selected for this chapter) works by creating random subsets of the whole transcription factor sample set, with a decision tree created from each subset. A ranking of classifiers is achieved by determining the number of times trees are split at a particular transcription factor. The regulation of each gene is treated as an individual subproblem, which is solved by employing feature selection to rank the interactions between all the (other) transcription factors and the target. Feature selection simplifies the model by taking into account only the most relevant features for network reconstruction.

GRENITS is based on a first-order linear autoregressive process, with modelling of uncertainty in the model and measurement process carried out within a Bayesian framework (Morrissey, 2013). Biological replicates are used in calculating the measurement error and improving the precision of the autoregression coefficients. This means that the true gene expression values are first inferred from the noisy data before the network topology can be retrieved.

Inferelator is an ordinary differential equation (ODE) based method for network inference with incorporation of prior knowledge, supplemented by time-lagged Context Likelihood of Relatedness (tlCLR) (Madar *et al.*, 2010). tlCLR is a mutual information metric similar to correlation which infers a directed relationship between two genes - it acts as a feature selection tool. The topology learned by tlCLR is then further improved by Inferelator, which employs least angle regression (LARS) to learn the network as a system of linear ODEs, all the while enforcing model sparsity.

TIGRESS treats the problem of regulatory network inference as a sparse regression problem, similarly to Inferelator and GENIE3, and employs least angle regression together with stability selection (Haury *et al.*, 2012).

2.2.1 Model ranking

Models inferred from the same data by different network inference algorithms can be very dissimilar, leading to a model selection problem. For instance, this difference between models trained on an Arabidopsis gene expression dataset can be seen in Table 2.8.

One method of discriminating between models is to parameterise the models and then determine which of the parameterised models explains the data best. This has been done using ODEs and Approximate Bayesian Computation to sample from the posterior distribution without calculating the likelihood (Csilléry *et al.*, 2010). This is advantageous as the likelihood - the probability of observing the data given a certain parameterised model - is often intractable. Approximate Bayesian Computation can be

improved using iterative refinement of parameters by incorporating sequential Monte Carlo (Toni *et al.*, 2009).

When the likelihood can be calculated, common model selection tools are the Bayes' Factor or Bayesian information criterion (BIC), which are closely related. The Bayes' Factor is used for comparing two models which are used to explain the same dataset by looking at the ratio of the marginal likelihoods:

$$BF_{12} = \frac{P(x|M_1)}{P(x|M_2)} \tag{2.1}$$

where $P(x|M_1)$ is the marginal probability of the data, x, under the model M_1:

$$P(x|M_1) = \int P(x|\theta, M_1)P(\theta)d\theta \tag{2.2}$$

where θ is the set of parameters in model M_1. The larger the Bayes' factor BF_{12}, the more strongly the data supports model M_1, as illustrated in Table 2.1.

The BIC score is defined as:

$$BIC_1 = -2\ln(P(x|M_1)) + k\ln(n), \tag{2.3}$$

where n is the number of data points of the data x, and k is the number of independent parameters. Given two models of the same parameter size, the connection between the BIC score and Bayes' Factor can be seen:

$$BIC_1 - BIC_2 = -2\ln(P(x|M_1)) + k\ln(n) - [-2\ln(P(x|M_2)) + k\ln(n)] \tag{2.4}$$

$$BIC_1 - BIC_2 = -2[\ln(P(x|M_1)) - \ln(P(x|M_2))]$$

$$BIC_1 - BIC_2 = -2\ln\frac{P(x|M_1)}{P(x|M_2)}$$

$$BIC_2 - BIC_1 = 2\ln BF_{12}$$

$$\Delta BIC_{21} = 2\ln BF_{12}$$

Hence the model with the lowest BIC score is more strongly supported by the data; the BIC score interpretation is illustrated in Table 2.1.

Table 2.1: Bayes Factor interpretation according to Kass & Raftery, 1995.

ΔBIC_{21}	Bayes Factor'	Evidence against M2
0 to 2	1 to 3	Weak
2 to 6	3 to 20	Substantial
6 to 10	20 to 150	Strong
>10	>150	Very Strong

The BIC and the related Aikake Information Criterion are often used in model selection for Bayesian networks (Liu *et al.*, 2012). BIC is a useful model ranking metric as it provides a balanced judgement of the goodness of fit between model and data, and model parsimony.

2.2.2 Employing GRNs for phenotype discovery

Forward or reverse genetic screens are often used to learn about the stress response, and its regulation, in plants (Luhua *et al.*, 2013). These methods involve random mutations (forward genetics), and knock-out/over-expression/knock-down of specific genes (reverse genetics). Forward genetics enables the discovery of novel genes through blind screening of pooled candidates for ones that perform a particular function. For instance, a T-DNA activation tagging library of 31,500 Arabidopsis lines was pooled, grown and screened for seed longevity under two conditions: natural and accelerated ageing (Bueso *et al.*, 2014). After an initial selection of 290 promising candidates, 4 mutants showed the desired seed longevity phenotype compared to wild type, and one was shown to have a mutation in a gene enoding a RING-type zinc finger putative ubiquitin ligase and further characterised (Bueso *et al.*, 2014). This is a success rate of just 0.01%, but clearly forward mutant screens are powerful when there is little understanding of a biological process and no obvious candidates for direct testing via reverse genetics. Other forward genetic screens show similar success rates (Kevei *et al.* (2006), Koiwa *et al.* (2006)).

Reverse genetic screens in Arabidopsis commonly use T-DNA insertion mutants from the Salk insertion lines ((Alonso *et al.*, 2003); `http://signal.salk.edu`) or Syngenta Arabidopsis Insertion Library (Sessions *et al.*, 2002), which are easily available via the Nottingham Arabidopsis Stock Centre (`http://arabidopsis.info/`) in the UK.

Equally, reverse genetic screens result in low rates of observation of the desired phenotype: $<$ 1-3% (Dobritsa *et al.* (2011), Pagnussat *et al.* (2005), Rama Devi *et al.* (2006)). However, when gene candidates are shortlisted using prior knowledge and experiments, the success rate is hugely increased. Luhua *et al.* (2013)) carried out phenotype screens on knock-out lines for 1007 genes previously identified to be differentially expressed during abiotic stress treatment. This involved screening these mutants for an altered stress response to osmotic, heat, cold, salinity, oxidative, and hypoxia stresses. 12-31% of the candidate genes screened showed tolerance or sensitivity, depending on the stress in question.

Thousands to hundreds of thousands of plants can be assessed in a typical high-throughput genetic screen. Given the low rate of positive results of these screens, it would be beneficial to find ways to prioritise candidates using network models - such as hubs or nodes of a certain hierarchical level - to lower the time and monetary costs involved. This has been successful in identifying genes that play a role in human disease. Several algorithms were used to combine network data with disease phenotype data, assuming that diseases with a similar phenotypic profile are likely to have functionally similar genes (Wu *et al.* (2008b), Li and Patra (2010)). In Arabidopsis, genes involved in certain processes have been identified through guilt-by-association using a network

integrating several types of omics data, AraNet (Lee *et al.*, 2010). Genes that are known to be involved with a trait are used as bait to discover other genes that play a role in that trait through their network connectivity. This guilt-by-association method has been effective at predicting genes participating in response to water deprivation, response to hydrogen peroxide, cold acclimation, response to heat, response to high light intensity and response to oxidative stress (Lee *et al.*, 2010). The construction of such regulatory networks to aid in screening candidates and minimise experimental work is therefore a promising approach, particularly given the availability of -omics datasets.

2.2.2.1 The link between network architecture and function in network models

As evidenced in the previous section, the main purpose of constructing biological networks is therefore to make testable predictions about biological systems (Lee *et al.*, 2010). For instance, in an update of the Arabidopsis circadian clock model, the incorporation of new data led to the insight that TOC1 is a repressor rather than an activator of the genes *LHY* and *CCA1*, which was confirmed by quantitative reverse transcription polymerase chain reaction (RT-qPCR) assays (Pokhilko *et al.*, 2012). Network properties of nodes have also been studied in order to obtain biological insight into gene function. Some common node and network characteristics are described in Table 2.2.

In protein-protein interaction (PPI) networks, it has been observed that deleting a hub, a protein with a large number of interaction partners, is far more likely to result in a lethal phenotype than deleting a non-hub (the centrality-lethality rule) (Jeong *et al.* (2001), He and Zhang (2006)). However, such rules have proven harder to decode for gene regulatory networks. An attractive and widely quoted theory is that biological networks (including GRNs) have certain architectural properties quite unlike random networks; they are scale-free, they have a hierarchical architecture, they exhibit an 'ultra small-world' effect, they are modular and they are enriched for certain network motifs (Barabasi and Oltvai, 2004). These properties are said to arise from the way biological networks have evolved, and thus the structure contains insight into the network's origins and function. The most often cited property is that biological networks are scale-free - this means that the degree distribution of nodes in the network follows a power-law, which arises from the preferential attachment of nodes and edges to existing hubs in the network. These scale-free networks are unusually robust to perturbations - which has been observed in biological networks (Isalan *et al.*, 2008) - but their hubs (the highly connected nodes) are very vulnerable. Because they are scale-free, the average path length of such a network is much smaller than expected in a random network. This allows fast communication as a node will be within a few links of most other nodes, due to the high connectivity of hubs (the ultra small-world effect). A unique property of scale-free networks is that subsets of the network have the same degree distribution and parameters as the whole network, allowing generalisations of the whole network to be made based on subnetworks; in other words, subnetworks of scale-free networks are themselves scale-free (Barabasi and Oltvai, 2004). A subnetwork is simply a subset of connected nodes from the original network.

Table 2.2: Common metrics used to describe networks and nodes. Most of these properties are properties of single nodes (degree, centrality, clustering coefficient); shortest path is property of two nodes and the degree distribution is a characteristic of the whole network.

Network characteristic	Description
Indegree	The number of incoming links to the node from other nodes in a directed network
Oudegree/hubbiness	The number of outgoing links from the node to other nodes in a directed network
Degree	The number of links of the node to other nodes. This is the number of incoming and outgoing links in a directed network
Clustering coefficient	A measure of the connectivity of the neighbours of the node: the number of links between neighbouring nodes divided by the total number of possible links between neighbouring nodes
Shortest path	The path with the smallest number of links between two nodes
Average path	The average of the shortest paths over all pairs of nodes
Betweenness centrality	The number of shortest paths passing through the node (sometimes normalised by dividing by the total number of shortest paths)
Pagerank centrality	Node importance score that takes into account the connections of the node and of its neighbours
Degree distribution	The probability distribution of the node degrees in the whole network

However, meta studies of published biological networks have disputed some of these claims, as they did not pass rigorous statistical tests (Lima-Mendez and van Helden (2009), Khanin and Wit (2006), Stumpf *et al.* (2005)). For instance, the degree distribution differs significantly from a power-law distribution according to a χ^2 goodness-of-fit test; rather it approaches a truncated power-law (Khanin and Wit, 2006). Therefore previously postulated evolutionary models are not enough to explain the appearance of biological networks (Rzhetsky and Gomez, 2001), and properties of a subnetwork cannot be used to draw conclusions about the properties of the unobserved parts of the network, as has been done before (Guelzim *et al.*, 2002).

The hierarchical features of gene regulatory networks are looked at in more detail in this chapter as it has received less focus in primary literature than other properties. Hierarchy refers to a network topology made of loosely interconnected network modules, with each module containing highly interconnected nodes (Ravasz and Barabási, 2003). One method of constructing a hierarchy for GRNs was introduced by Yu and Gerstein (2006). Using a breadth-first search, Yu and Gerstein (2006) have claimed that *Saccharomyces cerevisiae* and *Escherichia coli* GRNs have a pyramid-shaped hierarchical structure. This led to the observations that TFs in the top tier have more protein-protein interaction partners than TFs elsewhere in the hierarchy, that higher-level TFs influence more nodes, and that lower-level TFs are more essential to the organism. Therefore TFs at the top of the hierarchy are hypothesised to be signal inputs as they interact with more proteins on average and have the highest average closeness. Interestingly, the TFs at lower levels are more likely to produce lethal phenotypes when knocked out (Yu and Gerstein, 2006).

An additional refinement to the hierarchy construction was introduced by (Bhardwaj *et al.*, 2010). In order to avoid TFs at lower levels regulating TFs at higher levels, all the unassigned targets of an assigned TF were placed at its level. Their work was also conducted using *S. cerevisiae* and *E. coli* GRNs. They reached the opposite conclusion to Yu and Gerstein (2006): nodes at higher hierarchical levels are more essential to the organism. This is probably due to the fact that there was less observed redundancy in top tier TFs. These TFs were also expressed at a lower level and their mRNAs had a shorter half life.

Given the large amount of phenotypic data available for the response to *B. cinerea*, it would be interesting to determine whether the hierarchical position of a node affects the probability of its generating increased tolerance or sensitivity to *B. cinerea* when knocked out. Yu and Gerstein (2006)'s results could be due to the link between structure (high connectivity of top regulators) and function (ability to quickly influence many other nodes) as shown above. Perhaps higher tier Arabidopsis TFs able to influence more nodes are likely to have an impact on the infection outcome when knocked out. Additional network characteristics, such as betweeness centrality, clustering coefficient, time of differential expression, indegree, or outdegree, of a node may also help with the selection of those more likely to produce a phenotype when knocked out or overexpressed; the correlation between these and observed phenotype are also assessed.

2.3 Results

Five methods were ranked based on their performance on *in silico* and *in vitro* networks with known gold standards, as there is no unambiguous best way of inferring a gene regulatory network. The best algorithms (Marbach *et al.*, 2012) from different categories: regression (TIGRESS), mutual information (Inferelator), Random Forest (GENIE3), Gaussian process regression (CSI), and autoregression (GRENITS) were employed.

The returned networks from each algorithm consist of scores for edges starting at gene i to a gene j indicating the regulation of j by i, where the regulation can be either positive or negative. The score represents the confidence in the interaction. Self regulation cannot be inferred in any of these methods. This is regulation of a gene through a physical interaction from its own protein product. CSI avoids the issue of identifying self-regulation by assuming that all genes are affected by their own levels (Penfold and Wild, 2011). The other four methods do not attempt to look for self-regulation. Self-regulation is difficult to differentiate from the natural correlation of a gene's concentration at time t and the same gene's concentration at time $t+1$ or it can be a misinterpretation of RNA decay (Studham *et al.*, 2014). Furthermore, self-regulation is rare in biological networks (Vinh *et al.*, 2012). The most robust way of including self-regulation in the model is from experimental evidence such as yeast-1-hybrid assays.

The adjacency matrices, of size $p*p$, where p is the number of transcription factors in the network, are provided in Dataset F. The algorithms give a score for every single regulatory link possible. Therefore an important follow-up step for all these programs was the establishment of a threshold to limit the number of edges present in the network.

For each dataset, a consensus model of all 5 algorithms was also built using the AverageRank algorithm (Marbach *et al.*, 2012), as well as a consensus model between CSI and GRENITS (CSIGREN), and a consensus between GENIE3, Inferelator and TIGRESS (GENInfTIG). AverageRanks employs the Borda count method to rank all edges based on their scores from the given models. This method scores interactions from each network model so that the edges with the least confidence (the lowest score given by the network inference method) receive 1 point, the edge just above that receives 2 points and so on, with the top ranked interaction receiving the maximum number of points. This is done for all models, and the average score for edges across all models is calculated, and normalised (so that all edges receive a score from 0 to 1, with 1 indicating the most confidence in an edge) to form the edge scores for the consensus network. The top ranking edges were then kept in the consensus models.

Network models were constructed for TFs differentially expressed (DE) in Arabidopsis infected with *Botrytis cinerea*. Individual networks models were also constructed for DE TFs in Arabidopsis infected with *Pseudomonas syringae*, senescent Arabidopsis, Arabidopsis exposed to high light and drought-stressed Arabidopsis. These networks only contain Arabidopsis genes, and I refer to them as the *At-Botrytis* network, and so on, for simplicity's sake. These models are given in Dataset F. *In silico* networks and a murine network were first used for comparing the algorithms as there are very

little data available on what the 'true' Arabidopsis networks are.

2.3.1 DREAM network inference and model ranking

Schaffter *et al.* (2011) generated *in silico* network datasets of size 10 genes and 100 genes for the Dialogue for Reverse Engineering Assessments and Methods 4 (DREAM 4) competition for network inference methods using GeneNetWeaver (GNW) (Marbach *et al.*, 2009b). The datasets are publicly available and can be found at: `https://www.synapse.org/#!Synapse:syn3049712/files/`. There are 5 different transcription factor networks of size 10 nodes (DREAM 10.1, 10.2, 10.3, 10.4, 10.5). Each 10 gene network has 5 time series datasets, where a different perturbation has been applied to some of the nodes, with 21 time points in each dataset. The larger 5 DREAM networks with 100 genes have 10 datasets of 21 time points each. The time series data for the DREAM networks were generated using Stochastic Differential Equations (SDEs) to model internal noise in the processes of transcription, translation and degradation. There is also added measurement noise - a mix of normal and log-normal noise - similar to that found in microarrays. The perturbations are applied by modifying the basal transcription rate by a certain factor. Perturbations are added to a few random genes in the steady-state networks at time 0 and last for the first half of the time series. The perturbations are then removed for the second half.

The DREAM networks are suitable for testing inferred models and ranking methods on as they are subnetworks (the nodes and the connections between them) extracted from *E. coli* and *S. cerevisiae* (Marbach *et al.*, 2009a), and therefore biologically relevant. Perturbations applied to the genes can be seen as similar to the effects that stress such as infection has on Arabidopsis genes. In both cases, the perturbation/stress causes differential gene expression.

Models were inferred for all the DREAM networks by using the complete time series datasets as training data.

2.3.1.1 Area under ROC and PR curves

Networks were created from DREAM data as described previously. The predictions of the algorithms for the DREAM network models can be evaluated as the gold standard (or true network) is known. The inferred networks for DREAM datasets of size 10 nodes and size 100 nodes were used to calculate the receiver-operating characteristic area under curve (ROC AUC) and the precision-recall area under curve (PR AUC). The ROC curve plots the true positive ratio vs the false positive ratio of edges in the network at different network thresholds (see Equation 2.5). Each point on the ROC curve represents the true positive ratio and the false positive ratio at a particular network threshold.

The PR curve plots the precision against the true positive ratio, again for network models obtained at different thresholds. Precision compares the number of true positives to the number of true positives + false negatives, while the false positive ratio compares the number of false positives to the number of false positives and true negatives. In

datasets where there are a lot more true negatives than true positives, the false positive ratio can mask differences in performance due to the large number of true negatives. This problem does not occur when using precision (Davis and Goadrich, 2006).

$$\text{True positive ratio} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{2.5}$$

$$\text{False positive ratio} = \frac{\text{false positives}}{\text{false positives} + \text{true negatives}}$$

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

As a result, the PR curve is better at highlighting the difference between models in the case of the DREAM100 networks, where there are ∼200 true positives to ∼9700 true negatives. Taking this into account, the best DREAM10 models (according to the ROC AUC) are the GRENITS and consensus models. The best DREAM100 models (according to the PR AUC) are the CSI-GRENITS consensus and the overall consensus. In this case, the consensus network clearly performs best over the others according to the area under the ROC curve, while the CSI-GRENITS consensus performs best according to the area under the PR curve - see Tables 2.3 and 2.4.

Table 2.3: Area under the ROC curve for DREAM10 and DREAM100 networks. Area under the ROC curve for the 5 network models inferred from the individual network inference algorithms, the consensus between the 5 models, the consensus between CSI and GRENITS, and the consensus between GENIE3, Inferelator and TIGRESS. The model with the highest area is highlighted in green. The average AUROC is given at the end, together with the 95% confidence interval. The consensus model performs best overall.

	CSI	GENIE3	GRENITS	Inferelator	TIGRESS	consensus	CSIGREN	GENInfTIG
DREAM10.1	0.780	0.838	0.879	0.801	0.771	0.897	0.870	0.841
DREAM10.2	0.787	0.672	0.747	0.735	0.607	0.745	0.778	0.688
DREAM10.3	0.728	0.744	0.776	0.684	0.610	0.760	0.755	0.716
DREAM10.4	0.735	0.737	0.875	0.778	0.764	0.821	0.809	0.771
DREAM10.5	0.868	0.830	0.892	0.896	0.799	0.913	0.890	0.863
DREAM100.1	0.742	0.766	0.767	0.713	0.750	0.799	0.772	0.767
DREAM100.2	0.709	0.692	0.663	0.667	0.637	0.725	0.704	0.686
DREAM100.3	0.721	0.745	0.751	0.719	0.715	0.774	0.761	0.747
DREAM100.4	0.737	0.717	0.723	0.693	0.683	0.757	0.751	0.714
DREAM100.5	0.727	0.779	0.741	0.768	0.730	0.805	0.759	0.785
Average	0.753 ± 0.029	0.752 ± 0.033	0.781 ± 0.047	0.745 ± 0.042	0.707 ±0.043	0.800 ± 0.039	0.785 ± 0.035	0.758 ± 0.037

Table 2.4: Area under the PR curve for DREAM10 and DREAM100 networks. Area under the PR curve for the 5 network models inferred from the individual network inference algorithms, the consensus between the 5 models, the consensus between CSI and GRENITS, and the consensus between GENIE3, Inferelator and TIGRESS. The model with the highest area is highlighted in green. The average AUPR is given at the end, together with the 95% confidence interval. The consensus model between CSI and GRENITS performs best overall.

	CSI	GENIE3	GRENITS	Inferelator	TIGRESS	consensus	CSIGREN	GENInfTIG
DREAM10.1	0.470	0.396	0.602	0.336	0.342	0.611	0.644	0.382
DREAM10.2	0.439	0.258	0.450	0.315	0.216	0.397	0.533	0.259
DREAM10.3	0.399	0.327	0.533	0.237	0.197	0.398	0.588	0.251
DREAM10.4	0.524	0.273	0.730	0.344	0.265	0.600	0.688	0.315
DREAM10.5	0.402	0.398	0.620	0.451	0.320	0.769	0.699	0.436
DREAM100.1	0.118	0.062	0.139	0.050	0.058	0.136	0.151	0.063
DREAM100.2	0.109	0.058	0.080	0.056	0.046	0.086	0.117	0.057
DREAM100.3	0.111	0.082	0.107	0.077	0.077	0.130	0.122	0.085
DREAM100.4	0.109	0.063	0.120	0.059	0.057	0.114	0.135	0.066
DREAM100.5	0.113	0.076	0.099	0.078	0.074	0.132	0.122	0.084
Average	0.279 ± 0.111	0.199 ± 0.090	0.348 ± 0.162	0.200 ± 0.095	0.165 ± 0.072	0.337 ± 0.157	0.380 ± 0.166	0.200 ± 0.091

2.3.1.2 Bayesian Information Criterion score for model selection

A second way of evaluating the algorithms is by parameterising the different models, then using the BIC score to evaluate how well the models explain the parameterised data. Ideally, this method would separate bad models from good models as bad models would explain the data less well, even given the perfect set of parameters. The parameterisation is carried out for a long time so that the best possible parameters are inferred for each model, so that the goodness of fit relies only on the model itself. The BIC score is obtained from the maximal likelihood estimate of the data fitted using a Gaussian process (GP).

A minimisation function with 400 function evaluations is used to search for the hyperparameters that provide the best fit for the GP, given a certain network model (from the network inference algorithm) and training data (the DREAM time series data). The maximal likelihood can then be directly calculated given these hyperparameter values, the model and the training data. The maximal likelihood is then used to calculate the BIC score using Equation 2.3.

The advantage of using the BIC score is that it can rank models given the data and inferred parameters - it does not require a 'true network' to conclude that a model is superior to another. Therefore it can also be used to rank the Arabidopsis network models without referring to a gold standard as it judges how likely it is that the data came from a particular model. While the Bayes Factor can only be used to compare models in a pairwise fashion, a BIC score can be used to compare all models at once.

The use of the BIC score is evaluated on the DREAM10 and DREAM100 datasets, for which the true network, the gold standard, is known. Ultimately, the BIC score did not identify the gold standard as the best model most times, but rather the GRENITS model, therefore it was not used in evaluations of the Arabidopsis stress models (Table 2.5). At least 2 of the top 3 models as ranked by the PR and ROC AUC are also in the top 3 models as ranked by BIC.However, this method requires further refinement before using it for model selection.

2.3.2 Murine GRN model ranking

While the DREAM networks are useful for ranking algorithms, they are not as large as real-life networks, and are derived from prokaryotes rather than eukaryotes. For this reason, evaluation of the algorithms was carried out on an eukaryotic network for which a significant number of interactions are known. The input transcriptomic data is from murine liver (Zhang *et al.*, 2014). The transcriptome measurements were gathered every 2 hours for 48 hours. The gold standard was generated from ChIP-seq evidence for TF-promoter interactions for 9 circadian genes in the liver: *Bmal1, Clock, Npas2, Per1, Per2, Cry1, Cry2, Cbp, P300* (Koike *et al.*, 2012). Koike *et al.* (2012) located the DNA binding sites of these regulators every 4 hours for 24 hours. These data were then combined with known regulatory regions to produce a list of target genes. The number of targets for each regulator according to ChIP-seq experiment is shown in Table 2.6.

Table 2.5: BIC score for various models of the DREAM10 and DREAM100 networks. 'Gold' refers to the gold standard - the true network model. The best performing model is highlighted. The GRENITS models perform best overall.

	gold	CSI	GENIE3	GRENITS	Inferelator	TIGRESS	consensus	CSIGREN	GENInfTIG
DREAM10.1	291	318	304	286	328	315	311	296	316
DREAM10.2	343	338	357	319	348	356	336	320	348
DREAM10.3	329	340	340	318	343	343	320	324	332
DREAM10.4	311	335	330	311	313	338	316	328	326
DREAM10.5	261	254	275	256	286	271	249	258	266
DREAM100.1	2184	2195	2145	2129	2149	2163	2145	2161	2148
DREAM100.2	2954	2953	2921	2888	2911	2959	2905	2904	2912
DREAM100.3	2400	2399	2361	2361	2360	2414	2346	2366	2360
DREAM100.4	2562	2556	2529	2500	2520	2589	2518	2521	2524
DREAM100.5	2399	2390	2354	2319	2346	2411	2345	2358	2344

Table 2.6: Number of targets of mouse genes, as determined by ChIP-seq.

Gene	Number of targets
Bmal1	1040
Clock	869
Npas2	523
Per1	887
Per2	1191
Cry1	1789
Cry2	1477
Cbp	1433
P300	306

Network models were generated of all the rhythmical genes in the liver by selecting only genes with a rhythmical component in their expression (because only rhythmical genes were tested in the ChIP-seq data). The rhythmical component was determined by Zhang *et al.* (2014) using JTK-cycle evaluation, which is a programme that determines the periodicity and phase of transcripts from their temporal expression values (Hughes *et al.*, 2010). The expression profiles of 3189 genes was used for network inference, 356 of which were transcription factors. For the evaluation of the models, only the subnetwork consisting of the 9 regulators, together with all their possible targets (3189), was considered.

The generated network models with edge scores and the gold standard from the ChIP-seq were used to calculate ROC AUC and PR AUC - see Table 2.7. The model with the highest ROC AUC is the consensus. However, the GENIE3-Inferelator-TIGRESS consensus has the highest PR AUC. As the number of true positives is not dwarfed by the number of true negatives, the ROC AUC is preferred to the PR AUC.

A random network is formed by scrambling the edge scores of each network model. This is repeated 10 times, and used to calculate a mean ROC AUC and PR AUC for the random network model. The standard deviation for the random network model is also calculated and used to determine whether the inferred networks perform significantly better than random (are two standard deviations away). Only the GENIE3, Inferelator, TIGRESS, GENInfTIG and overall consensus perform better than a random network model. This is quite unlike the results obtained for the DREAM network, where GRENITS and the CSI-GRENITS consensus performed very well overall.

2.3.3 *At-Botrytis* Transcription Factor Networks

The consensus network model was overall the best model from the previous sections, but not conclusively so. It was the most commonly ranked top model in terms of area under the ROC curve, and consistently near the top in terms of area under the PR curve. While the CSI-GRENITS model performed very well in terms of the area under the PR curve for the DREAM networks, it was not significantly better than using

Table 2.7: Area under the ROC and PR curves for mouse gene regulatory network models inferred from the individual network inference algorithms, the consensus between the 5 models, the consensus between CSI and GRENITS, and the consensus between GENIE3, Inferelator and TIGRESS. The top performing model is highlighted in green.

	CSI	GENIE3	GRENITS	Inferelator	TIGRESS	consensus	CSIGREN	GENInfTIG
AUROC	0.525	0.519	0.532	0.527	0.514	0.534	0.532	0.520
AUPR	0.358	0.359	0.351	0.363	0.363	0.361	0.354	0.365

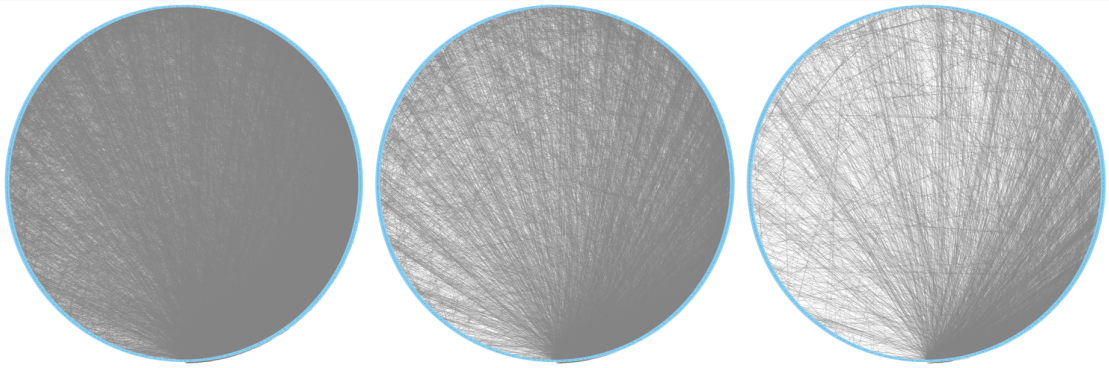


Figure 2.1: Models from the *At-Botrytis* consensus network thresholded to have 17660, 8830 and 4415 edges, from left to right respectively. There are 883 nodes in the model. The visualisation was created in Cytoscape using the degree-sorted circle layout, which positions nodes in order of degree around a circle, with the node of highest degree positioned at 6 o'clock.

a random network for the murine model. The model performance seem very much to be dependant on the data used, and much more reliable on small networks.

In view of this, the process of using all 5 algorithms to infer a network model was repeated for the dataset of Arabidopsis infected with *B. cinerea*. In this time series, leaf 7 from four different plants were infected with *B. cinerea* and sampled at each time point every 2 hours, with sampling beginning at 24 to 48 hours post infection (hpi). At the same time, 4 uninfected leaves were also sampled for all time points, to provide controls for their infected counterparts. Differentially expressed genes (DEG) were identified by the original authors Windram *et al.* (2012) using a combination of the F test within MAANOVA (Wu *et al.*, 2003) and a Gaussian process two-sample test (Stegle *et al.*, 2010). 5706 genes were identified as being downregulated and 4112 as being upregulated. This was narrowed down to a TF-only dataset, with 386 upregulated genes and 497 downregulated genes, for a total of 883 nodes. Only TFs were selected for this analysis in order to reduce the computational time and the size of the final network. Additionally, only TFs will be considered for modelling purposes in subsequent chapters, as they form complex interactions (non-TF genes will not have an impact on other genes in the network) and their impact on the network when manipulated is greater.

A threshold for the *At-Botrytis* models was chosen to create a sparser network, and retain the edges with the highest confidence. The scores produced by each algorithm are not directly comparable, so thresholding was performed so that each model had the same number of edges. All networks were initially thresholded so that the total number of edges were 5x, 10x or 20x the total number of nodes $\approx$ 4415, 8830 and 17660 respectively. Visualisation of the consensus network models at these three thresholds are given in Figures 2.1 and 2.2.

A very low number (167) of edges are common to all 5 of these algorithm models at this particular network size (10x). However, the models returned by CSI and GREN-

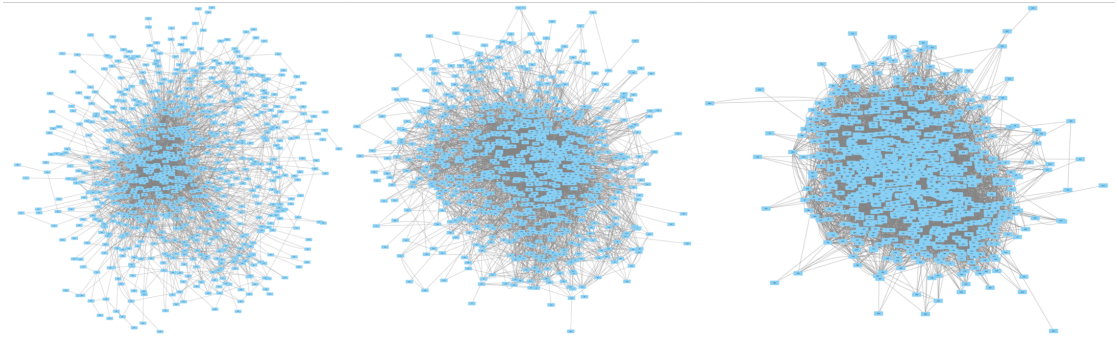


Figure 2.2: Models from the *At-Botrytis* consensus network thresholded to have 4415, 8830 and 17660 edges, from left to right respectively. There are 883 nodes in the model. The visualisation was created in Cytoscape using the spring-embedded layout, which contains nodes of higher degree in the centre, and nodes with lower degree on the outside.

ITS have a high similarity, as do the models returned by GENIE3, Inferelator and TIGRESS - see Table 2.8. This is probably a factor of the mathematics underlying these algorithms - CSI relies on Gaussian process regression and Bayesian inference, GRENITS is a Bayesian inference method and GENIE3, Inferelator and TIGRESS are feature selection/regression-based. This similarity was also observed in the earlier DREAM and murine networks (although it is not shown).

Different stringencies were tested in the network models, to see if more stringent thresholds resulted in a higher proportion of overlap between the different algorithms, and hence a higher confidence network. This is based on the assumption that higher confidence edges are more likely to be conserved across algorithm models than lower confidence edges. However, it seems that the number of overlaps between the GENIE3, Inferelator and TIGRESS networks is directly proportional to size, and the same can be said of the CSI and GRENITS networks - see Figure 2.3.

Table 2.8: Common edges between *At-Botrytis* network models. Each model is thresholded so that it has ≈ 17660 edges. (a) shows the number of edges common to all pairwise combinations of models. (b) shows the percentage of edges common to all pairwise combinations of models.

a

	CSI	GENIE3	GRENITS	Inferelator	TIGRESS
CSI					
GENIE3	1583				
GRENITS	4551	1673			
Inferelator	929	7204	1161		
TIGRESS	968	7171	1203	8264	

b

	CSI	GENIE3	GRENITS	Inferelator	TIGRESS
CSI					
GENIE3	9				
GRENITS	26	9			
Inferelator	5	41	7		
TIGRESS	5	41	7	47	

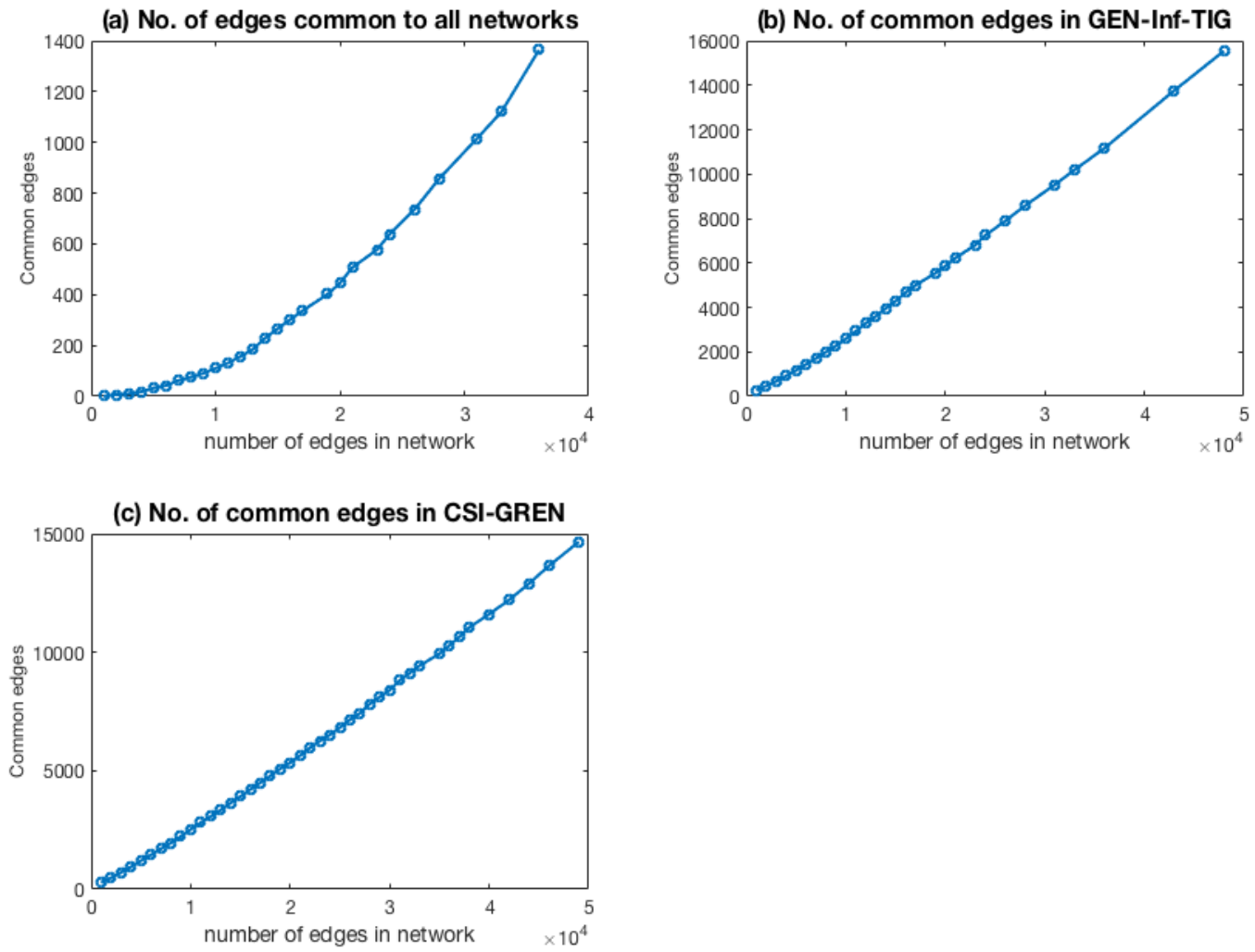


Figure 2.3: Common edges between *At-Botrytis* network models. The thresholds for each model are varied so the number of common edges for different network sizes can be determined. (a) Number of edges found in all 5 networks vs. network size. (b) Number of edges found in GENIE3, Inferelator and TIGRESS models vs network size (c) Number of edges found in CSI and GRENITS model vs. network size.

The topology of the different models was also examined. The degree distribution - indegree, outdegree, and total number of edges for each node - was plotted in figures 2.4-2.6 for each network model of size 4415 edges. For an explanation for each of these network characteristics, refer to Table 2.2. This reveals some interesting differences in the degree distribution between models. Regarding the indegree - the number of regulators for each node - nodes in all the models have 25 or fewer. However, there is a difference between the mode (the value that occurs most frequently) number of regulators - with clear peaks at 1 for GENIE3, 5 for CSI, GRENITS and TIGRESS and a flatter peak at 1-5 regulators for Inferelator (Figure 2.4). CSI, GENIE3 and GRENITS make predictions for larger hubs in the network with 100 or more targets, whereas Inferelator are more conservative with their estimate of <30 targets regulated by the largest hubs (Figure 2.5). While the consensus, GENIE3-Inferelator-TIGRESS consensus and GENIE3 histograms appear to follow a power law distribution (Figure 2.6), these are actually truncated power-law distributions (data not shown).

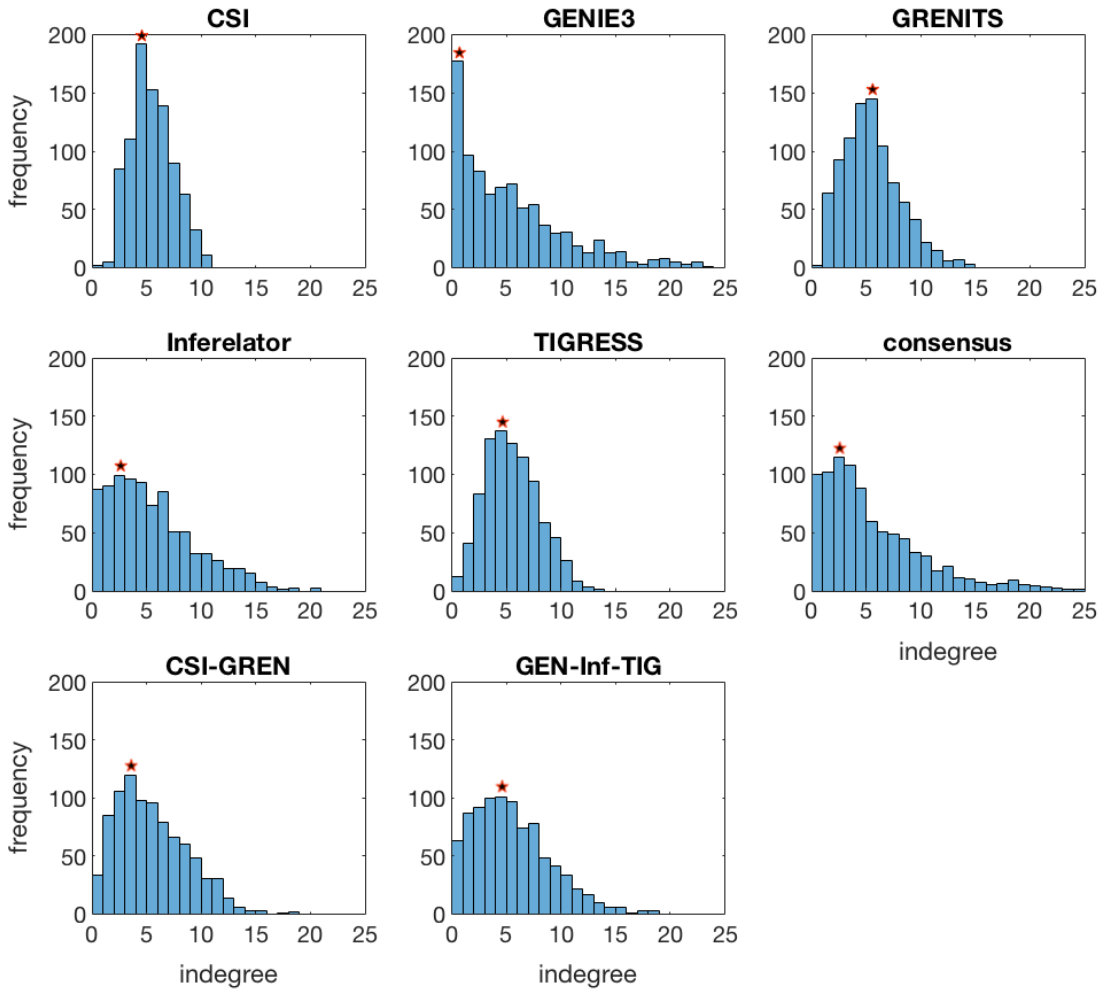


Figure 2.4: Histogram of model indegree for nodes of the *At-Botrytis* networks of 4415 edges. The nodes in all models have a maximum of 25 regulators. Each bin has width 1. The mode is identified by a star.

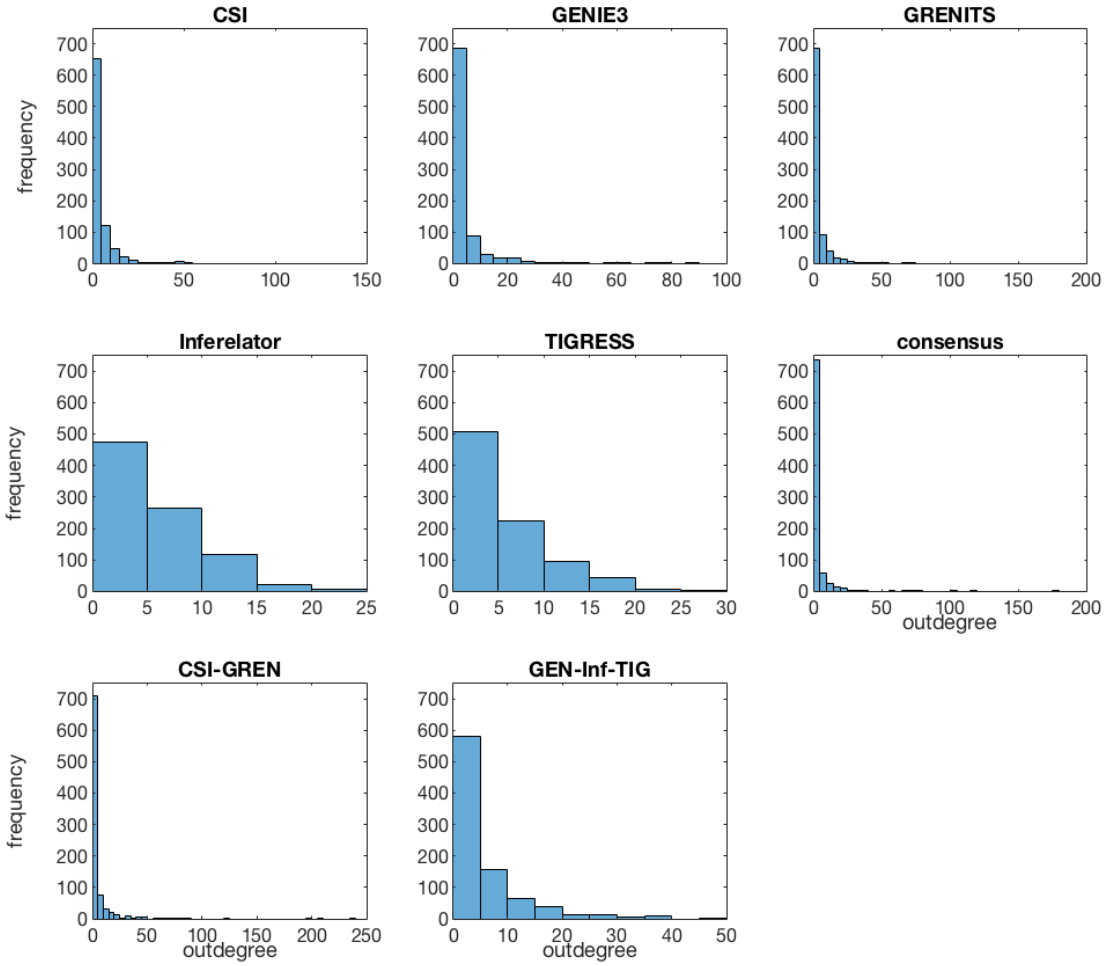


Figure 2.5: Histogram of model outdegree for nodes of the *At-Botrytis* network of 4415 edges. The maximum outdegree varies widely for each model from 25 (Inferelator) to 188 (GRENITS). Each bin has width 5.

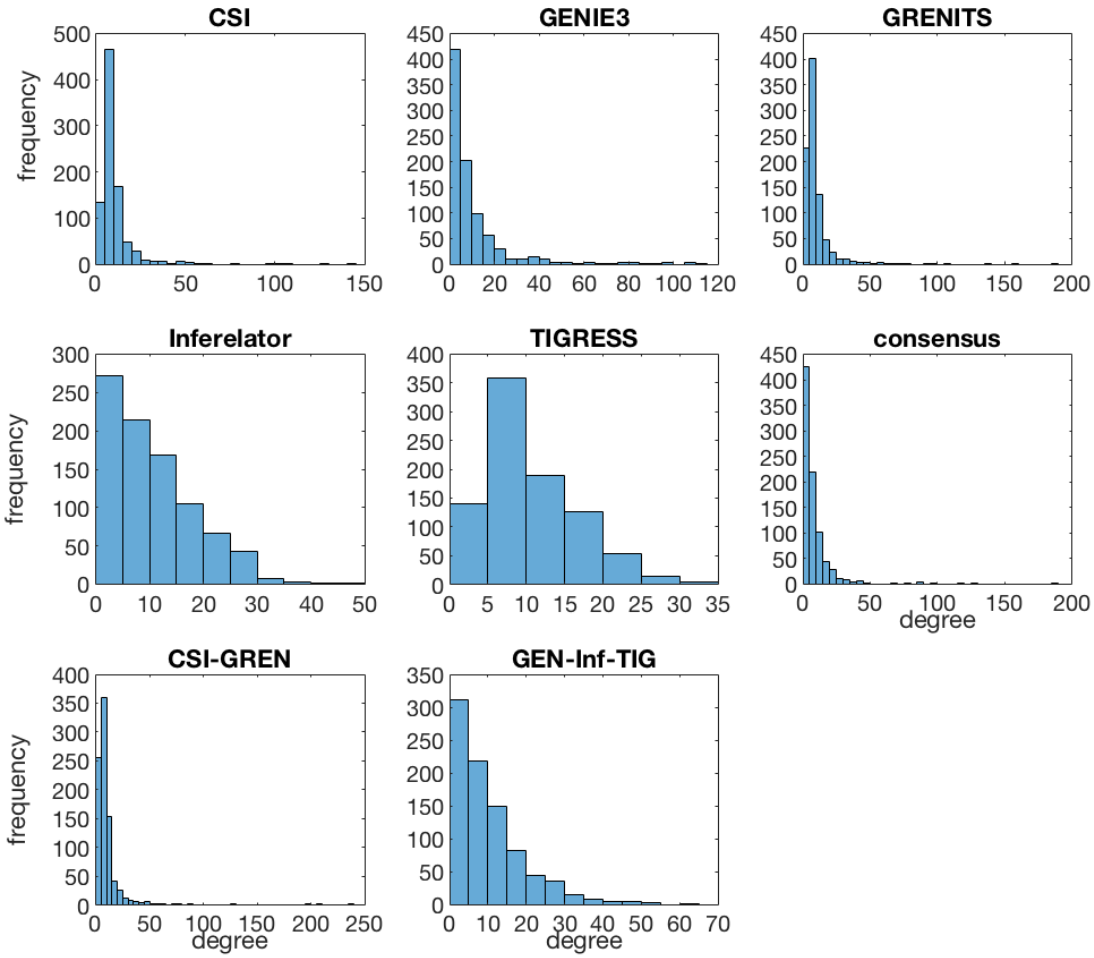


Figure 2.6: Histogram of model degree (indegrees + outdegrees) for nodes of the *At-Botrytis* networks of 4415 edges. Each bin has width 5.

2.3.3.1 Biological data for model ranking

Following network inference, experimental data was used to rank the resulting models. Biological data such as Yeast-1-Hybrid (Y1H), ChIP-Seq, DNA affinity purification sequencing (DAP-Seq) (O'Malley *et al.*, 2016), and gene expression levels after knock-out (KO) and over-expression (OE) of a TF can be used to validate edges in network models, and enhance our confidence in the models by being incorporated as priors. Y1H, ChIP-Seq and DAP-Seq data directly measure the ability of proteins to bind to regulatory DNA regions. KO and OE studies can point to direct or indirect regulatory links between TFs and downstream targets, but there is no way to distinguish between the two without further experiments.

A total of 970 TF-target interactions between the 883 Arabidopsis TFs were obtained from Y1H data from unpublished experiments performed during the Warwick

University PRESTA project, and the ATRM (Jin *et al.*, 2015) and AtRegNet databases (Yilmaz *et al.*, 2011). ATRM contains curated networks through data mining of 974 peer-reviewed papers. AtRegNet collects evidence for direct interactions between TFs and promoters from Y1H, electrophoretic mobility shift assay and chromatin immunoprecipitation. It also takes into account interactions inferred through glucorticoid-mediated transcriptional induction systems (Aoyama and Chua, 1997) and supported by further data from KO/OE studies. These 970 experimentally-verified edges are used to rank the *At-Botrytis* network models.

As mentioned in Section 2.3.1.1 on the DREAM network model rankings, for datasets with a small percentage of true positives compared to true negatives, it is preferred to use PR AUC over ROC AUC. However, online databases such as ATRM (Jin *et al.*, 2015) and AtRegNet (Yilmaz *et al.*, 2011) containing Y1H and ChIP-Seq datasets typically report only the positive results. Therefore there is no way of knowing all the regulatory links that were tested and inferring the true negatives. Additionally, without the true negatives, there is no way of calculating the false positives in the network as there are still untested interactions. As a result, the number of true positives (edges present in the model and the 970 biologically-verified interactions) is plotted against the network size in Figure 2.7. The number of true positives in a network of size 17660 edges is also presented in Table 2.9.

From this data, it can be seen that the networks inferred by GENIE3, TIGRESS and Inferelator are superior for this purpose to the ones inferred by GRENITS and CSI; the consensus model is ranked in between these two groups (Figure 2.7.

	consensus	CSI	GENIE3	GRENITS	Inferelator	TIGRESS	CSI-GREN	GEN-Inf-TIG
No. of TP	37	25	51	28	66	49	18	58

Table 2.9: Y1H and ChIP-chip/seq data-verified edges in a network of size = 20x the total number of nodes ≈ 17660 edges for consensus, CSI, GENIE3, GRENITS, Inferelator, TIGRESS, CSI-GRENITS consensus and GENIE3-Inferelator-TIGRESS consensus networks using known direct transcription factor-DNA interactions from Y1H and chromatin immunoprecipitation experiments.

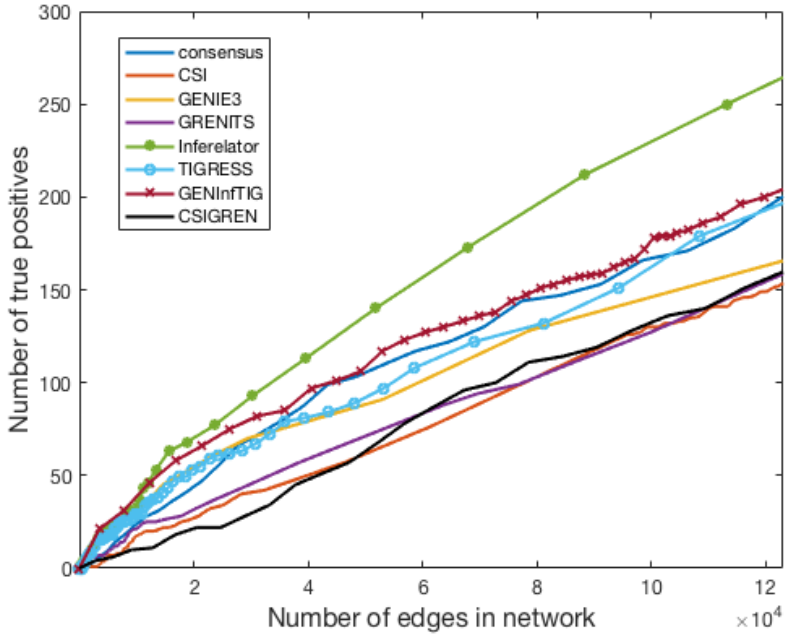


Figure 2.7: Y1H and ChIP-chip/seq data-verified edges in various network sizes for consensus, CSI, GENIE3, GRENITS, Inferelator and TIGRESS networks. Experimental data consists of known direct transcription factor-DNA interactions from databases containing Y1H and chromatin immunoprecipitation experiments.

Transcriptomic measurements from 10 unpublished PRESTA experiments were also used to verify the accuracy of the edges in the *At-Botrytis* models. These transgenic plants contain a particular TF knocked out using T-DNA inserts (Krysan *et al.*, 1999) or

overexpressed using a 35S promoter (Kay *et al.*, 1987). These plants were then infected with *B. cinerea* and their gene expression levels measured at one or more time points using microarrays (see Table 2.16). The list of differentially expressed genes cause by the KO/OE was obtained by comparing gene expression in the KO/OE mutant during *B. cinerea* infection with gene expression in a wild type Arabidopsis Col-0 during *B. cinerea* infection. A gene was considered differentially expressed if it had a fold change of 2 (either up- or downregulated) and an adjusted (for multiple testing) p-value < 0.05.

If a knockout/overexpression of transcription factor X causes a gene to be differentially expressed, then the assumption is that this gene must either be directly regulated by TF X, or indirectly regulated, through intermediary transcription factors or by other means. The downstream DEGs were assumed to be at most 3 degrees removed from the KO/OE TF X in the network models (see Figure 2.8). A true positive was counted if a DEG in the KO/OE experiment of TF X was also present downstream (maximum of three degrees downstream) of that TF X in the *in silico* network models. A false negative occurred when a DEG in the KO/OE experiment of TF X was not present in the neighbourhood of TF X in the *in silico* network.

The number of true positives was then plotted against different network sizes. At a small network size - $< 20,000$ edges - GRENITS and CSI perform best when looking at true positives amongst first- and second-degree neighbours, whereas Inferelator performs better at larger network sizes (Figure 2.9 and Table 2.10). When first-, second- and third-degree neighbours are taken into account, CSI, Inferelator and GRENITS models perform best for small and large networks. The consensus model has middling performance.

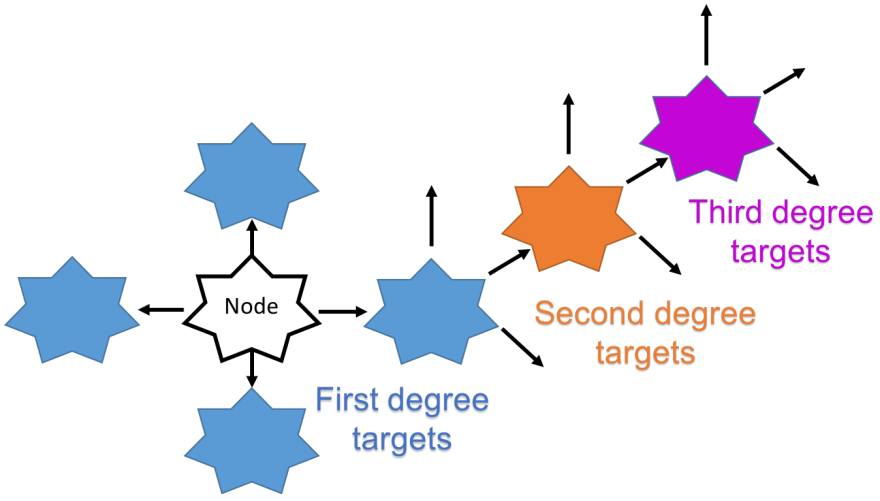


Figure 2.8: First-, second- and third-degree neighbours of a node. TFs such as those in knockout/overexpression experiments can affect target genes directly or indirectly. In this figure, nodes are TFs and the nodes in blue are direct or first degree targets and the nodes in orange and pink are indirect targets of the original node.

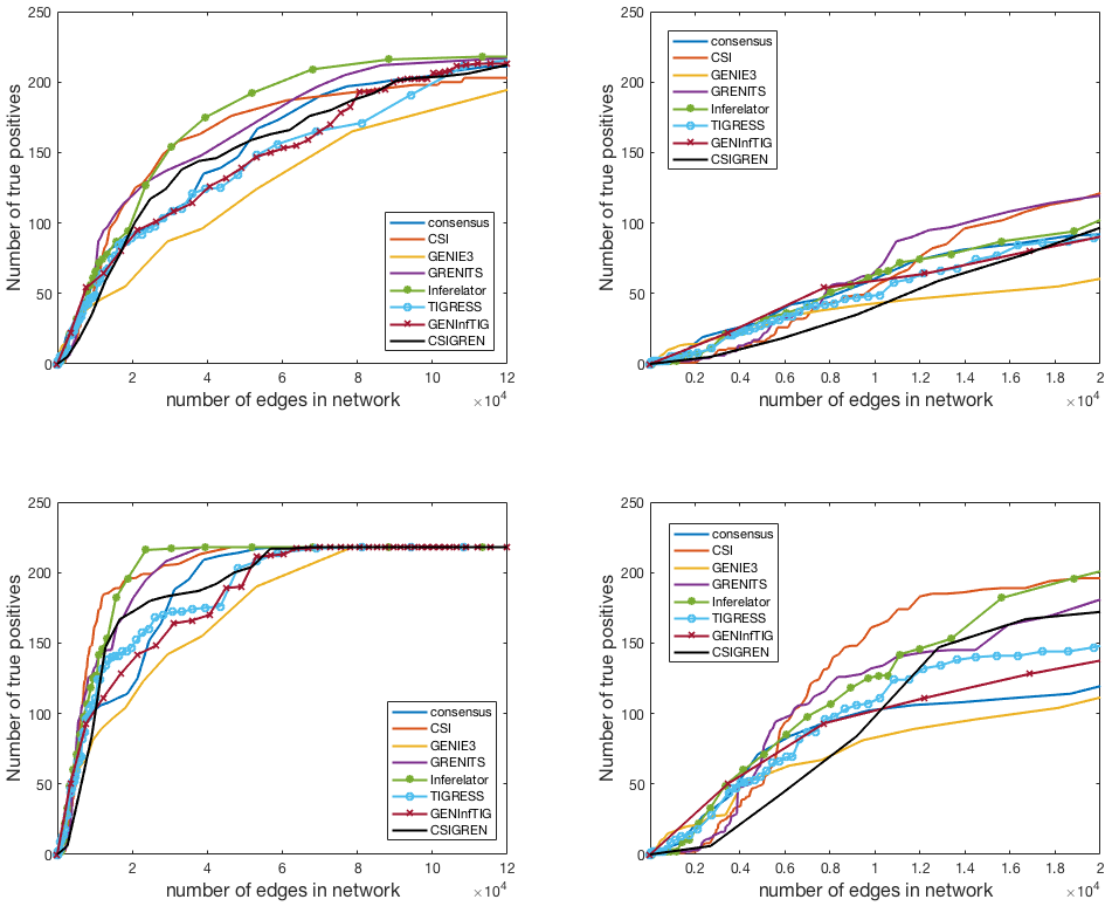


Figure 2.9: KO and OE data-verified edges in consensus, CSI, GENIE3, GRENITS, Inferelator, TIGRESS, CSI-GRENITS consensus and GENIE3-Inferelator-TIGRESS consensus network models of various sizes. The top 2 figures only take into account first- and second-degree targets; the bottom two figures take into account first-, second- and third-degree targets. This is based on the TFs that are differentially expressed in an Arabidopsis knockout or overexpressor during *B. cinerea* infection.

It is surprising that, given how different the models are (Table 2.8), they had similar number of true positives when tested with both kinds of biological data - they seemingly pick up different sets of true interactions.

Whilst the various network ranking tools applied did not agree on a single algorithm or model as being clearly superior, the consensus network has always performed well (direct and indirect regulatory data for the *At-Botrytis* network) or better than average (mouse GRN, DREAM GRNs). The literature also suggests that consensus models are more robust (Marbach *et al.* (2012), Marbach *et al.* (2010)), therefore I have decided to carry out simulations and further work on the consensus model over the other models.

Table 2.10: Number of KO and OE data-verified edges during *B. cinerea* infection in the *At-Botrytis* models. Number of true positives in a network of size = 20x the total number of nodes ≈ 17660 edges for consensus, CSI, GENIE3, GRENITS, Inferelator, TIGRESS, CSI-GRENITS consensus and GENIE3-Inferelator-TIGRESS consensus networks.The top table only takes into account first- and second-degree targets; the bottom table takes into account first-, second- and third-degree targets.

	consensus	CSI	GENIE3	GRENITS	Inferelator	TIGRESS	CSI-GREN	GEN-Inf-TIG
No. of TP	90	113	54	114	92	86	81	84

	consensus	CSI	GENIE3	GRENITS	Inferelator	TIGRESS	CSI-GREN	GEN-Inf-TIG
No. of TP	113	194	104	169	193	144	169	133

2.3.4 *At-Botrytis* consensus network structure and function

2.3.4.1 Node characteristics and function

Networks can be used to make predictions about the functional role of a particular TF (node) based on network properties, which reduces the number of experiments that need to be performed. For instance, in Özgür *et al.* (2008), data mining was used to build a network between 226 genes, 17.7% of which were related to prostate cancer. They then generated lists of top ranked genes using centrality network measures. Degree and eigenvector centrality were able to identify 95% of the genes that are known to play a role in prostate cancer within their top ranked 20 genes. A similar approach to (Özgür *et al.*, 2008) was implemented by using network properties to rank genes with and without associated *B. cinerea*-susceptibility phenotypes.

A literature search was carried out to identify Arabidopsis genes that are positive or negative regulators of the defence response to *B. cinerea*. The experiments in question used knockouts/overexpressors infected with *B. cinerea* to identify the genes from transgenic plants that have modified resistance compared to wild type Arabidopsis. This was combined with further such unpublished data from the PRESTA project. Genes that produce more susceptible plants when KO, or more resistant plants when OE are positive regulators of defence. Genes that produce more resistant plants when KO, or more susceptible plants when OE are negative regulators of defence. Out of the nodes in the 883-gene network, 50 are negative/positive regulators of defence, and 52 do not appear to play an observable role in the disease process, or are redundant.

A number of network features were evaluated to see whether there is a difference in features between TFs that are involved in defence and TFs that are not (as determined by KO/OE studies - lack of a phenotype when a TF is KO/OE does not necessarily mean it is not involved in defence). These features include the indegree, outdegree, betweeness centrality, time of differential expression, clustering coefficient, and average change in gene expression values of TF nodes. The node characteristics were determined for the *At-Botrytis* consensus network. This information can further be used to reduce the network model to a smaller number of key nodes in order to simplify modelling approaches.

Spearman rank-order correlation (for evaluating relationships between ordinal variables) showed that none of the chosen network characteristics significantly discriminate for positive/negative regulators of defence. The mean, together with error bars, of the metrics, is shown in Figure 2.10 for TFs that do and do not have an impact on defence. However, a positive correlation has previously been made between the number of regulators of a gene and the strength of its expression (Hansen and O'shea (2013), Heyndrickx *et al.* (2014)). There is a significant correlation in the *At-Botrytis* network between the indegree of a gene and its expression amplitude with a p value <0.001 after Bonferroni correction for multiple correlation tests (Figure 2.11). This may be because individual TFs have only a small influence on the mRNA production rate, whereas multiple (positive) regulators would greatly increase transcript levels.

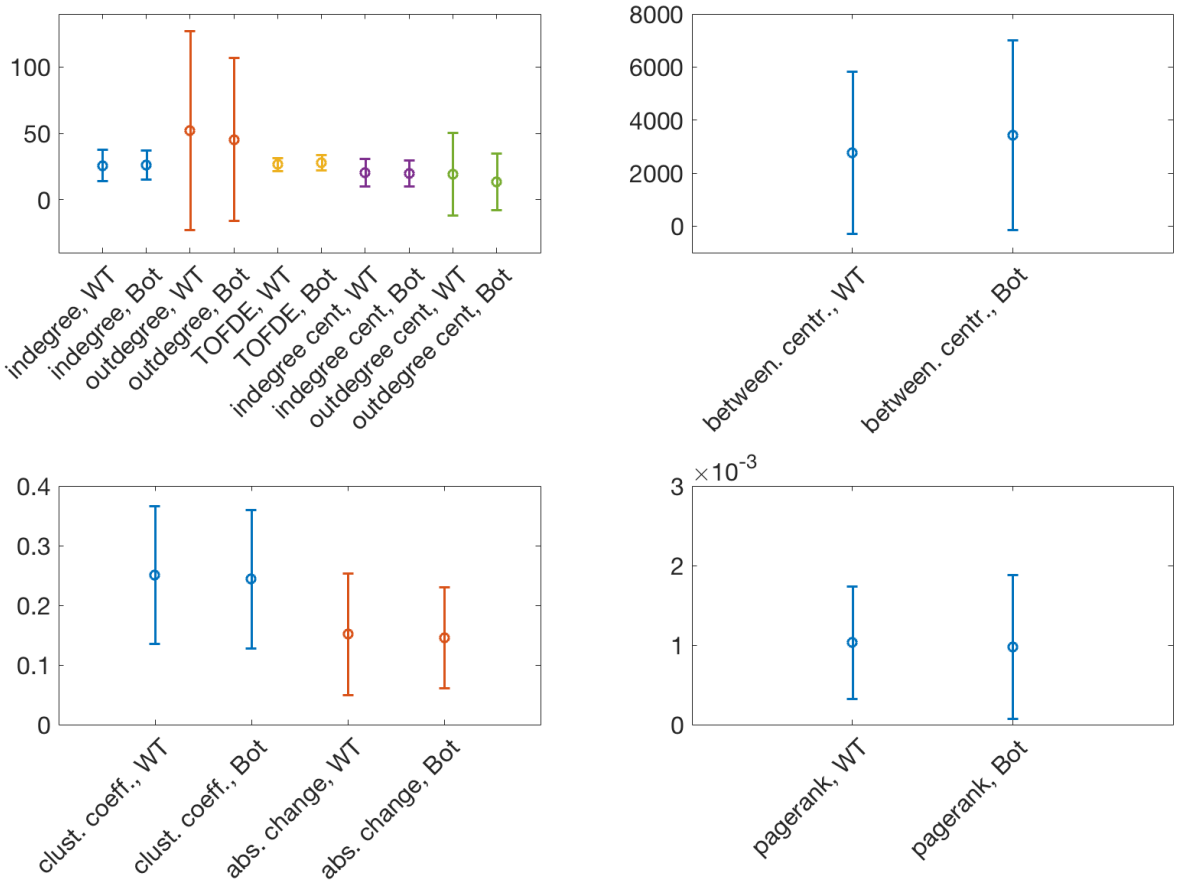


Figure 2.10: Network characteristics for TFs which are positive/negative regulators of defence and TFs which do not influence the defence response when KO/OE. Mean (circles) and standard deviation (error bars) for network metrics of nodes that do not impact defence (labelled 'WT') and nodes that are positive/negative regulators of defence (labelled 'Bot') in the *At-Botrytis* consensus network. The network characteristics assayed are indegree, outdegree, time of differential expression in *At-Botrytis* timeseries dataset (first plot), betweeness centrality (second plot), clustering coeffcient and absolute change in expression levels in *At-Botrytis* timeseries dataset (third plot), and pagerank centrality (fourth plot).

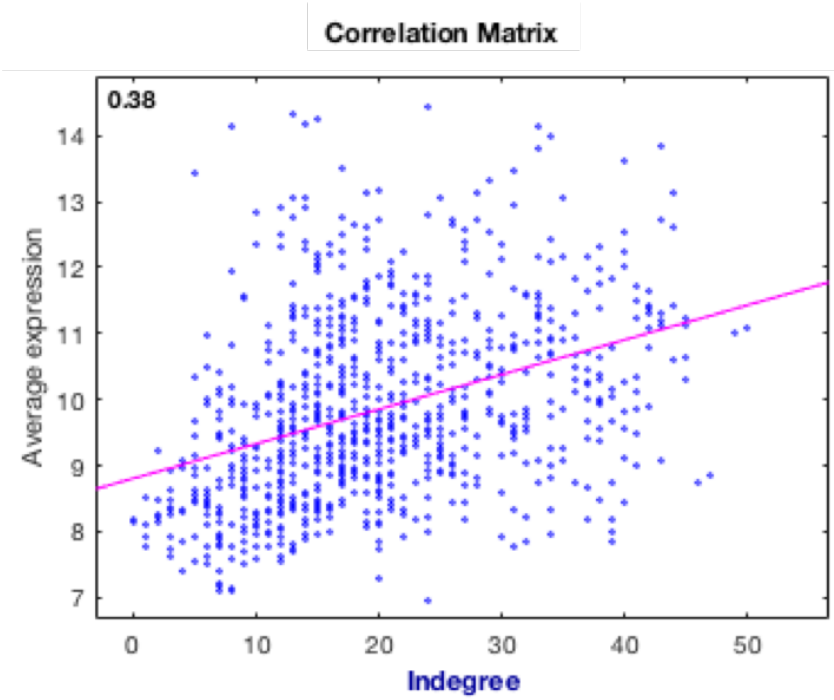


Figure 2.11: Correlation coefficient for node indegree and average expression is significantly different from zero (p-value <0.001). Scatter plot of average gene expression vs. indegree. Line of best fit is shown in pink; the correlation coefficient is 0.38. Analysis was performed for the consensus *At-Botrytis* network thresholded at 17660 edges.

2.3.4.2 Hierarchical network structure and function

The hierarchy of nodes in a network can also be an indicator for node function. A hierarchy consists of a pyramidal structure of nodes, where very few nodes are present in the top tier, and these control the nodes in the level below, which control the nodes in the level below them. Lower tiers are more populated with nodes than upper tiers. This structure is similar to that seen in governmental and corporate organisations, with departmental managers supervising middle managers, who manage regular workers. Based on the methods employed by Yu and Gerstein (2006), a hierarchy was built of the *At-Botrytis* consensus model. The transcription factors that did not regulate other TFs were assigned to the bottom of the hierarchy - Level 0. The transcription factors that regulated the TFs at the bottom of the hierarchy were assigned to level 1 of the hierarchy, the transcription factors that regulated the TFs on level 1 of the hierarchy were assigned to level 2, and so on. This method can also potentially help with reducing the number of nodes (currently 883) by identifying a particular hierarchical level(s) that contains most of the interesting nodes. Two network sizes were used in the building of hierarchies: networks of either 17760 (20 times the number of nodes) or 4415 edges (5 times the number of nodes). The hierarchical model contains 6 or 7 levels, depending

on the cut-off point for number of edges.

The number of genes at each hierarchical level, as well as the number of genes at each level that do and do not affect the defence response are shown in Figures 2.12 and 2.13. Hypergeometric tests were carried out for each level to determine whether there is enrichment for defence genes. The hypergeometric test, rather than a Student t-test, is chosen as it calculates the statistical significance of having x successes out of y total successes + failures, when there is no replacement - therefore it is the distribution of choice for identifying over and under-represented sub-populations in a sample. There is no significant enrichment at any hierarchical level for TFs that result in altered susceptibility to *B. cinerea* when knocked out (compared to what would be expected if these TFs were randomly distributed among hierarchical levels; Table 2.11).

Yu and Gerstein (2006) observed in their networks that TFs in the top tier were more likely to participate in protein-protein interactions (PPI). They hypothesised that this is because top-tier TFs are global modulators that integrate signals from various stresses and pass the information on through their regulatory links. The number of Arabidopsis TF PPIs and its correlation with hierarchy level was tested in the *At-Botrytis* network. A PPI map was available from Trigg *et al.* (2017) of 8577 interactions among 1453 TFs. The number of interactions that each TF in the *At-Botrytis* network had was extracted from this data. For the TFs that did have at least one protein-protein interaction, their hierarchical level was also recorded. The average number of PPIs for each hierarchical level was then calculated for both network sizes (see Table 2.12). While there does seem to be a larger average of PPIs in level 4 of the hierarchy for the network of 17760 edges, this effect of larger average PPI for top tier TFs is not seen in the smaller network size. Therefore, this is not a reproducible observation, but may be just an artefact seen at some arbitrary thresholds.

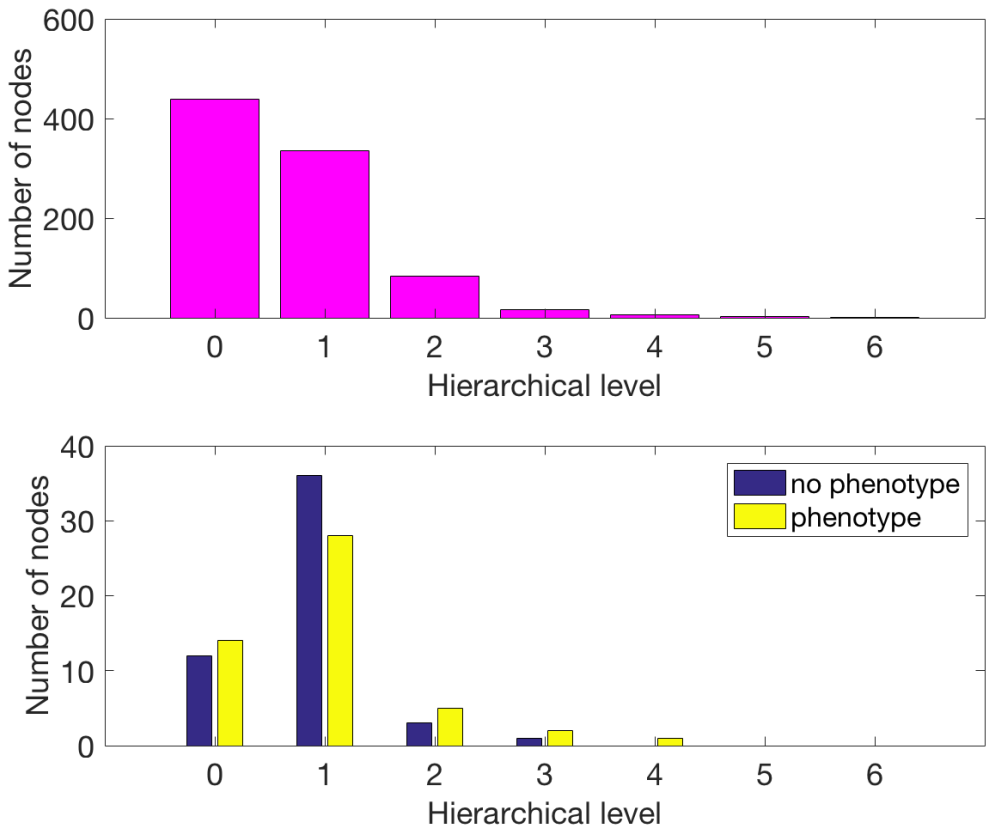


Figure 2.12: The number of nodes at each hierarchical level in the *At-Botrytis* consensus network of 4415 edges (top), and the number of nodes that do and do not influence the *B. cinerea* disease process at each hierarchical level (bottom). P-values are shown in Table 2.11.

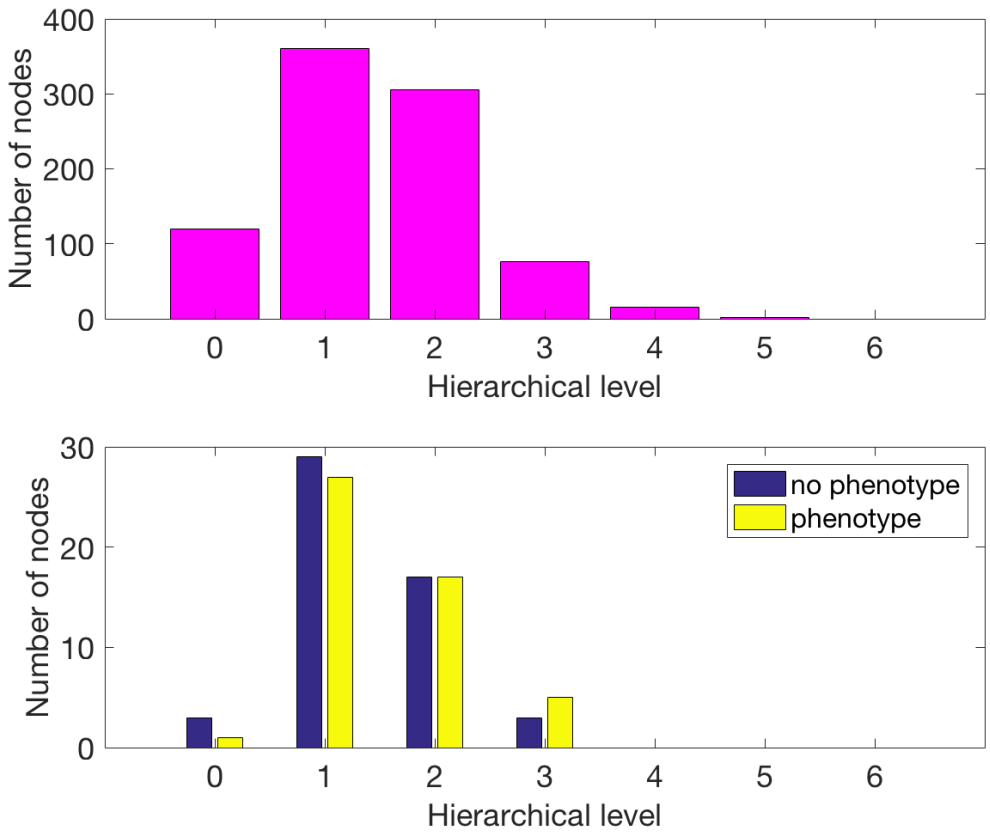


Figure 2.13: The number of nodes at each hierarchical level in the *At-Botrytis* consensus network 17660 edges (top), and the number of nodes do and do not influence the *B. cinerea* disease process at each hierarchical level (bottom). P-values are shown in Table 2.11.

Alternatively, breadth-first search can first be employed as in Yu and Gerstein (2006) to construct hierarchies, and additionally loops can be treated as in Bhardwaj *et al.* (2010). This produces a different hierarchical structure to the one described previously, and may carry a higher correlation between hierarchy level and TF function. This results in a hierarchy with a reduced number of levels from the added specification that all targets of a TF cannot belong to a level higher than that TF. However, this hierarchy too does not produce additional insight into the question of phenotype discovery (Figure 2.14). This method of hierarchy construction resulted in a very flat hierarchy with only 3 levels (albeit with very few nodes at level 2) unlike in the original study (Bhardwaj *et al.*, 2010), where *E. coli* and *S. cerevisiae* networks were found to have 4 hierarchical levels. In addition, the lower gene expression at higher hierarchical levels as seen in Bhardwaj *et al.* (2010) was not observed in this study.

Table 2.11: P-values for over- and under-enrichment of nodes that do and do not influence the disease process at the different hierarchical levels based on the hypergeometric distribution. The null hypothesis is that there is no under- or over-enriched at any level. The total sample size is the number of experimentally tested TFs (102). The number of successes is the number of TFs affecting the disease process when KO/OE; the number of failures is the number of TFs that do not affect disease when KO/OE.

20x network

Level	Sample size	Phenotype (successes)	P value
0	4	1	0.32
1	56	27	0.51
2	34	17	0.53
3	8	5	0.34
4	0	0	-
5	0	0	-

5x network

Level	Sample size	Phenotype (successes)	P value
0	26	14	0.37
1	64	28	0.12
2	8	5	0.34
3	3	2	0.49
4	1	1	0.49
5	0	0	-
6	0	0	-

Table 2.12: Average number of protein-protein interactions for TFs belonging to each level in the network hierarchy produced using the method from Yu and Gerstein (2006). This was calculated for networks of two size: 4415 and 17760 edges. The number of genes at each level is also shown.

Level	No. genes, 17760 network	PPI, 17660 network	No. genes, 4415 network	PPI, 4415 network
Level 0	120	9.05	439	10.64
Level 1	361	11.05	336	12.22
Level 2	306	11.23	83	9.80
Level 3	76	10.11	16	7.63
Level 4	16	39.57	6	7.00
Level 5	2	2.00	2	-
Level 6	0	-	1	-

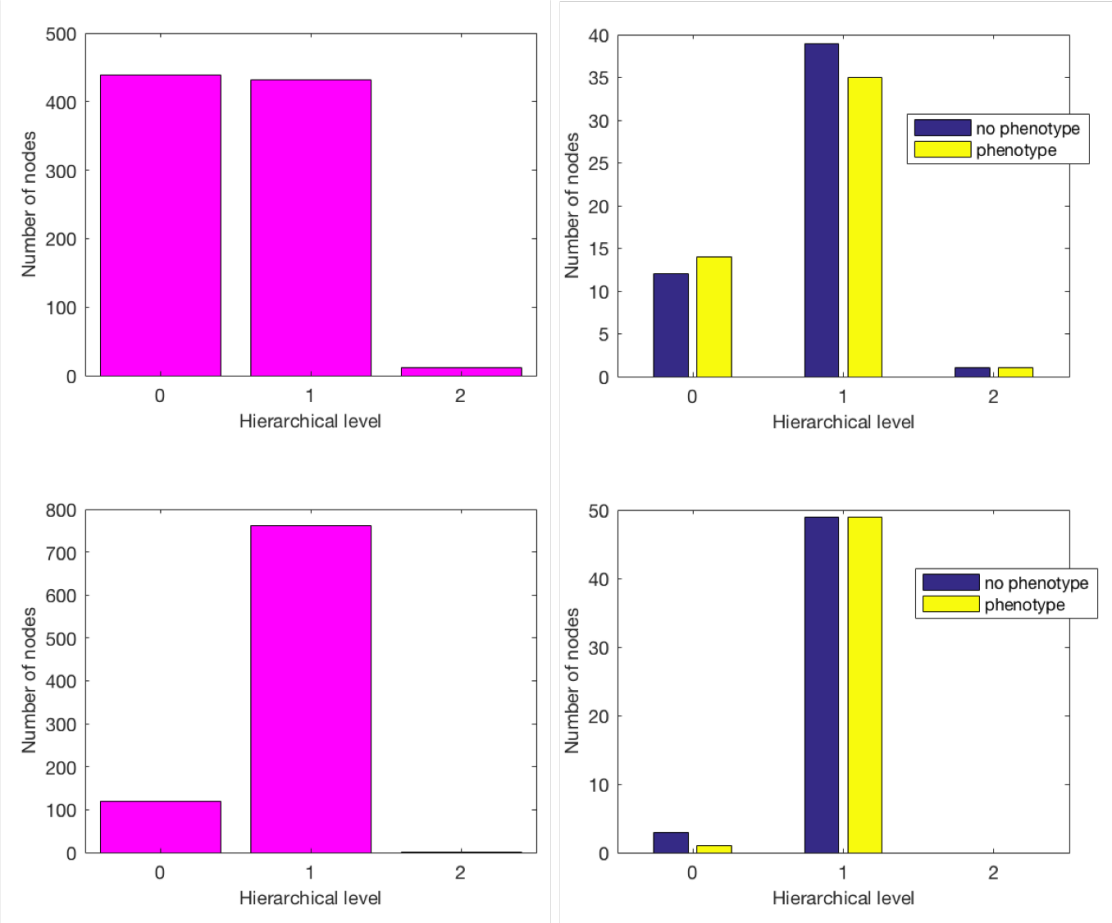


Figure 2.14: Number of nodes in the hierarchy of the *At-Botrytis* consensus network with 4415 edges (top row) and 17660 edges (bottom row) constructed according to (Bhardwaj *et al.*, 2010). Number of genes with and without an impact on defence upon KO/OE are shown in the plots on the right.

Two different methods for creating hierarchies and studying connections between hierarchical structure and node function were employed in this work (Yu and Gerstein (2006), Bhardwaj *et al.* (2010)). Neither method of organisation revealed that biologically important nodes aggregate at a particular level. Given that their published hierarchies were constructed based on *E. coli* and *S. cerevisiae* networks, it may be that eukaryotic networks do not exhibit such features. Their networks were constructed from biochemical, genetic and ChIP-chip data. This could present a bias whereby nodes that are targeted by other nodes, but whose regulatory targets have not been experimentally determined, are placed in the lowest tier. In addition to this, confirmation bias may also exist - TFs whose regulatory interactions have already been studied using ChIP or Y1H are more likely to be selected for PPI screening, compared to other TFs about which little is known.

2.3.5 A Multiple Stress Arabidopsis Transcription Factor Network

Given the availability of time series transcriptomic data for Arabidopsis during drought, high light, long day senescence, short day senescence and infection with *P. syringae* and *P. syringae hrpA*, it would be interesting to explore the intersection of TF networks constructed from different stresses. Information from multiple stresses can also be used for reducing the size of screening populations, as shown in Ransbotyn *et al.* (2015). Ransbotyn *et al.* (2015) were able to achieve a phenotype screening success rate of 62% using their network model constructed from abiotic stress data. To do this, they used microarray data corresponding to 10 abiotic stress treatments, and gave each gene a score based on the repeatability of its differential expression in the different experiments. The top scoring genes were then used to generate a co-expression network, out of which several candidate genes were chosen for screening in salt, osmotic and heat stress. The resulting success rate was 10-fold greater compared to simply screening random candidates. Here, I aim to construct a similar network with genes that are differentially expressed in biotic and abiotic stresses; however, this network is directed rather than the coexpression network of Ransbotyn *et al.* (2015). Spearman rank correlation did not produce very accurate networks as determined by Marbach *et al.* (2012).

First, a consensus network model was inferred for each condition. The input data once again consisted of the expression levels of DE TFs and was used to infer directed network models with the 5 algorithms as discussed in the section above, which were then combined into a consensus using AverageRank. These consensus models were thresholded to contain a number of edges equivalent to 20 times the number of nodes. A common Arabidopsis stress network was generated, containing nodes and edges found in multiple stresses. This was done by including all the edges that were present in at least 2 networks, and their corresponding nodes. However, this did not include edges that were only present in the two *Pseudomonas syringae* datasets or only the two senescence datasets. All the networks, after thresholding, were pairwise compared in order to determine where the stress responses overlapped. The final network had 797 nodes and 1799 edges. Only 7 TFs were DE in all the stress networks.

The 6 additional stress conditions were: *Pseudomonas syringae*-infected WT and

hrpA- Arabidopsis mutant, short day senescence and long day senescence, drought and high light stresses. The inferred TF network models are all given in Dataset F. The *At-Botrytis* network contains 883 TFs, the *At-Pseudomonas* WT network 879 TFs, the *At-Pseudomonas hrpA*- network 562 TFs, the *At*-long day senescence network 575 TFs, the *At*-short day senescence network 520 TFs, the *At*-drought network 288 TFs and the *At*-high light network 587 TFs.

Despite having a significant number of overlapping nodes, not many regulatory interactions were actually present between any two different stress networks (Table 2.13). However, regulatory interactions in this network should be of higher confidence (compared to interactions in each individual stress network), as they have been predicted independently in at least 2 cases.

The genes in the common stress network were ranked according to the number of TFs they regulated, and the top 20 hubs are given in Table 2.14. Out of the top 20 hubs, 13 have already been identified as playing a role in the Arabidopsis stress response. AT2G28200, AT4G14920, AT1G64530, AT3G05545, TIFY1, BBX14 and BLH7 have not yet been found to play a role in the Arabidopsis response to stress.

4 of the hubs are NAC TFs. This is one of the largest families of TFs specific to plants, and they have been found to play roles in development, biotic and biotic stress (Olsen *et al.*, 2005). ANAC046 is a positive regulator of senescence associated genes and chlorophyll catabolism (Oda-Yamamizo *et al.*, 2016). ANAC013 facilitates the communication between mitochondria and the nucleus upon mitochondrial stress, leading to oxidative stress tolerance (De Clercq *et al.*, 2013). ANAC047 is most well-known for regulating the biosynthesis of ethylene in response to flooding in order to promote growth and raise the level of leaves so that they remain in contact with air (Rauf *et al.* (2013), Hofmann (2013)). ATAF1 (ANAC002) is a negative regulator of the response to abiotic stresses (salt, and oxidative stress) and the response to necrotrophs (Wu *et al.*, 2009b). 3 of the hubs are NF-YAs. NF-YA4 forms a complex with bZIP28 which upregulates endoplasmic reticulum stress-related genes (Liu and Howell, 2010). NF-YA10 is thought to play a positive role in the response to abiotic stress (Leyva-González *et al.*, 2012). NF-YA1 regulates growth under salt stress (Li *et al.*, 2013). Two of the hubs are WRKY TFs. The WRKY family of transcription factors is one of the largest, with 74 members involved in biotic and abiotic stresses, seed development, and senescence, among other processes (Rushton *et al.*, 2010). WRKY45 is a positive regulator of phosphate uptake in Arabidopsis and is a key player during phosphate starvation, by upregulating the phosphate transporter PHT1;1 (Wang *et al.*, 2014). WRKY48 is weakens the Arabidopsis defence response to *P. syringae*, probably by downregulating (directly or indirectly) defence-related PR genes (Xing *et al.*, 2008). OCP3 mediates the Arabidopsis resistance to necrotrophs by jasmonic acid signalling through COI1 (Coego *et al.*, 2005) and is also involved in drought tolerance (Ramírez *et al.*, 2009). ANT is primarily seen to regulate flower development, however, it has also been shown to regulate defence pathways, and an *ant6 ail6* double knockout showed increased levels of salicylic acid and jasmonic acid, and increased resistance to *P. syringae* (Krizek *et al.*, 2016). ABF1 is thought to play a role in the drought stress response and is also shown to be differentially expressed upon

Table 2.13: Pairwise comparison of common DE TFs found in 7 Arabidopsis stress consensus networks (top table). The total number of DE TFs for each stress is given at the bottom. Pairwise comparison of common edges found in 7 Arabidopsis stress consensus networks (bottom table). The total number of edges in each stress network is given at the bottom.

	Botrytis	Pseud WT	Pseud hrpA	Long day senes	Short day senes	Drought	Highlight
Botrytis							
Pseud WT	471						
Pseud hrpA	327	461					
Long day senes	307	352	269				
Short day senes	279	285	193	320			
Drought	162	146	104	122	116		
Highlight	317	292	190	215	192	122	
Total no. DE TF	883	879	562	575	520	288	587

	Botrytis	Pseud WT	Pseud hrpA	Long day senes	Short day senes	Drought	Highlight
Botrytis							
Pseud WT	144						
Pseud hrpA	110	243					
Long day senes	103	74	90				
Short day senes	68	64	49	198			
Drought	60	50	39	45	56		
Highlight	60	81	59	64	45	113	
Total no. edges	17660	17580	11240	11500	10400	5760	11740

Table 2.14: The top 20 hubs in the common Arabidopsis stress network.

AGI	Name
AT3G04060	ANAC046
AT5G11270	OCP3
AT4G24470	TIFY1
AT1G68520	BBX14
AT3G01970	WRKY45
AT2G28200	C2H2-type zinc finger family protein
AT2G34720	NF-YA4
AT4G37750	ANT
AT4G14920	Acyl-CoA N-acyltransferase with RING/FYVE/PHD-type zinc finger protein
AT5G12840	NF-YA1
AT1G01720	ATAF1
AT2G16400	BLH7
AT5G06510	NF-YA10
AT1G32870	ANAC013
AT1G49720	ABF1
AT5G49520	WRKY48
AT1G64530	Plant regulator RWP-RK family protein
AT3G05545	RING/U-box superfamily protein
AT2G27050	EIL1
AT3G04070	ANAC047

drought and high salt stresses and ABA treatment (Yoshida *et al.*, 2015). EIL1 is a key player in ethylene and jasmonic acid signalling, as detailed in Section 1.3.3.

The genes in the various individual networks, including the common stress network, were then ranked according to their outdegree, or hubbiness. This was done with a view to identify TFs that are hubs in multiple networks. 7 top-20 hub genes were found in more than one stress network. These are presented in Table 2.15. Two of these are also top hubs in the common stress network.

Table 2.15: Top hubs found in more than 1 Arabidopsis stress network

Top hub	ATI	*At* stress network
WRKY65	AT1G29280	Long day senescence + short day senescence
NF-YA5	AT1G54160	B. cinerea + drought
BBX14	AT1G68520	High light + hrpA + common stress
WRKY66	AT1G80590	Pseudomonas + hrpA
NF-YA2	AT3G05690	Short day senescence + drought
TIFY1	AT4G24470	B. cinerea + drought + common stress

WRKY66 has been hypothesised to interact with group A, B and C MAP kinases (Pathak *et al.*, 2013). It is induced by salicylic acid (Besseau *et al.*, 2012). However, its response has not been studied in depth in *Pseudomonas* infection, making it a potential candidate for further work. WRKY65 is differentially expressed during carbon starvation (Contento *et al.*, 2004), and has also been hypothesised to interact with group A MAP kinases and Group B MAP kinases, which mediate signal transduction for biotic and abiotic stresses (Pathak *et al.*, 2013). However, its function in senescence is uncharacterised. The involvement of NF-YA5 in drought is known - NF-YA5 overexpressors are more drought-tolerant than WT and NF-YA5 knockouts are less tolerant (Li *et al.*, 2008). The expression levels of NF-YA5 increases during drought, and that is due to the downregulation of *miR169a*, which targets the NF-YA5 transcript for cleavage (Li *et al.*, 2008). The induction of NF-YA5 is abscisic acid-dependant. NF-YA5 is also two-fold upregulated in response to high salinity and 40-fold upregulated in high sucrose (Leyva-González *et al.*, 2012). NF-YA5 has shown to be downregulated during the *B. cinerea* stress response in tomato, as a result of upregulation of *miR169* (Jin *et al.*, 2012). NF-YA2 is also involved in response to biotic stress and a target of the *miR169* family (Leyva-González *et al.*, 2012). Its levels are increased in response to low phosphorus, low nitrogen, high sucrose and low phosphate availability (Leyva-González *et al.*, 2012). Overexpression of this TF causes a dwarf phenotype but also increases the tolerance to nitrogen deprivation: WT plants had clear senescent symptoms, while the NF-YA2 OE did not (Leyva-González *et al.*, 2012). This perhaps ties in with its prominence as a hub in the *At*-short day senescence network. BBX14 belongs to the B-Box Zinc-Finger family of TFs. As the name suggests, they contain both a Zinc finger and a B-box motif and are thought to bind both DNA and protein (Khanna *et al.*, 2009). Its expression is downregulated in *ala6* knockouts, a gene that maintain membrane stability during high temperatures (Niu *et al.*, 2017) and it is upregulated by cytokinin treatment (Balazadeh *et al.*, 2014). TIFY1 is a TF containing a ZIM domain and GATA-type zinc-finger domain, and its overexpression causes hypocotyls and petioles to be elongated (Shikata *et al.*, 2004). The ZIM domain is also found in JAZ (Jasmonate ZIM-domain) proteins, which repress jasmonate signalling. TIFY1's role in the stress response has not been studied in depth, however, the ZIM domain is involved in protein-protein interactions and the formation of homo and heterodimers (Chung and Howe, 2009). There is a possibility that TIFY1 interacts with JAZ family members via their ZIM domains and influences the response to stress this way.

These TFs, together with the top hubs in the common stress network make a promising candidate list for phenotype assays to various stresses. Coming up as hubs in multiple networks may be an indication that these genes play an important role in defence.

2.4 Discussion

This chapter inferred networks for Arabidopsis transcription factors that were differentially expressed in response to biotic and abiotic stresses. The data used to build

the networks were transcriptomic time series. These models were then used to explore the relationship between network structure and gene function as well as the node and edge similarities between stresses.

Of the network inference algorithms, two were based on Bayesian inference (CSI, GRENITS), and three employed variations on feature selection (GENIE3, Inferelator, TIGRESS). The models produced by similar algorithms showed a much greater similarity. Consensus networks were also created between CSI and GRENITS, between GENIE3, Inferelator and TIGRESS, and between all five models. Despite doing a number of performance evaluations using *in silico* networks, murine TF networks and Arabidopsis experimental results, no one model was strikingly better. As has been seen before, different algorithms perform better in different settings and there is no particular category of inference methods that is superior to others (Marbach *et al.*, 2012). However, as also observed by Marbach *et al.* (2012), the consensus of all five network inference algorithms had the most robust performance. Therefore the *At-Botrytis* consensus model was chosen to carry further analysis out on. Consensus models were similarly constructed for Arabidopsis TFs responding to other perturbations: *P. syringae* infection, drought, high light, and senescence. The networks inferred in this work are unique as models of this size have not been assembled for Arabidopsis TFs responding to stresses.

A promising method of ranking network models based on the available data, rather than the true network structure, is with the Bayesian Information Criterion. This is useful as the true network structure is not always available. For example, in the case of the *At-Botrytis* network, (partial) KO/OE data is only available for 10 TFs, while less reliable Y1H and ChIP-seq data is only available for a subset of TFs. The BIC score is based on the maximal likelihood derived from a parameterised Gaussian process model. However, the BIC score failed to identify the gold standard as the best model when tested out on DREAM10 and DREAM100 networks. This is perhaps because the BIC assumes that the hyperparameter values inferred are the 'true' hyperparameter values, and this is not necessarily so; the minimisation function used to derive the hyperparameters may have reached a local rather than global optimum. Using the point estimate - the maximal likelihood - of the hyperparameters is also not as desirable as integrating over all hyperparameter values for the purpose of model selection.

Some Arabidopsis experimental data was available which was used to differentiate between the models produced by different algorithms (Section 2.3.3.1). In particular, the most readily available TF-binding site data for Arabidopsis was from Y1H experiments. However, these are carried out in yeast, in conditions completely different from those found in Arabidopsis targeted by *B. cinerea*. Therefore lack of an edge in the model from Y1H studies does not necessarily mean the interaction does not occur *in planta*. Binding of a TF to its target may be context dependent - it may only occur during infection. ChIP-seq data is often seen as the gold standard for evaluating the performance of network inference algorithms (Desai *et al.*, 2017; Kulkarni *et al.*, 2017). However, these experiments are usually carried out in non-stressed Arabidopsis, and binding of a TF to a promoter region does not necessarily mean regulation of the downstream gene. Only a quarter of all binding effects are found to have any consequence on the expression of the

target gene (Swift and Coruzzi, 2017). False negatives may also arise as TF regulation does not always occur through binding proximal to the target gene. ChIP datasets were only available for a few Arabidopsis TF as they are more labour intensive and expensive to carry out than Y1H.

While transcriptomic data from KO/OE experiments carried out in Arabidopsis infected with *B. cinerea* are a more reliable source for true positives, these are highly specific and not many such datasets are available. It is also impossible to differentiate between direct and indirect regulations without further experiments or data. Expression data from KO/OE experiments will also not reveal regulatory edges masked by redundancy. An additional cause for concern is the small overlap in genes of ChIP targets and DEG in KO or OE studies of the same TF (Swift and Coruzzi (2017), Figure 3). This disparity between technique results could be due to the large false negative rate of ChIP experiments (Chen *et al.*, 2012), transient binding not captured by ChIP or Y1H (Swift and Coruzzi, 2017), functional redundancy of the TF, TF localisation in the nucleus being signal dependant (García *et al.*, 2010; Kaminaka *et al.*, 2006), and Y1H false positives and negatives (Reece-Hoyes and Walhout, 2012).

It is possible to use this same experimental data instead to increase our confidence in a GRN. These data provide direct evidence that a transcription factor binds a certain DNA sequence and are equivalent to the presence of an edge in the network. Combining both GRNs and experimental data can overcome the weaknesses inherent in both methods: the interactions found with Y1H and ChIP-seq may not be relevant to the infection process, and the edges found with network inference may not be physically possible *in vivo* due to the lack of binding sites. As a result, the final 'best' network is the consensus combined with experimentally-confirmed interactions using AverageRank: the Y1H results, the interactions from AtRegNet and the ATRM. The direct TF-promoter binding data from Y1H and ChIP studies is not taken to be an edge with strict certainty as this does not necessarily demonstrate regulation between the TF and its target; rather it is used as a prior that strengthens our confidence in certain edges (Section 2.5.7). This is the network that will be used for simulation purposes in further chapters. The results from KO/OE studies were not included as priors in this final network as they are only available for 10 genes in the network - this would bias the final network to edges from these genes.

However, even with the incorporation of these experimental data into GRNs, network inference remains a challenging problem for a variety of reasons. It is impossible to say how accurate these networks are due to our lack of knowledge of the true underlying network. The number of data points is still small compared to the number of nodes in the network, making inference challenging. This is known as the high dimensionality problem and it makes it impossible to select between different models that explain the observed data equally well (Banf and Rhee, 2017).

It is also possible that TFs vital to the stress response have been discarded by the initial screening for not being differentially expressed. However this does not necessarily mean they play no role in the infection process (Quackenbush, 2001). The algorithms used assume that similar - but delayed - expression patterns indicate that a gene is

being targeted by the TF - however that is not necessarily the case. There could be a delay between regulation and a change in gene expression. On the other hand, the gene expression measurements may have been taken at too large time intervals, therefore missing important biological events. Additionally, these algorithms are also prone to inferring more feedforward loops than actually exist. This is because a simple linear pathway such as X $\rightarrow$ Y $\rightarrow$ Z can easily be interpreted to have a link X $\rightarrow$ Z and the algorithms cannot distinguish between direct and indirect regulation.

An improvement upon modelling the Arabidopsis defence response with GRNs would be the use of multiscale models which incorporate data from different sources such as gene, protein and metabolite levels, to mechanical responses at the cell, tissue and organ level. By integrating data from different scales, a more holistic picture of the disease process can be obtained. Multiscale models have been recently used to understand the relationship between genotype and phenotype in Arabidopsis roots (Band *et al.*, 2012; Hill *et al.*, 2013) and understanding host-microbe interactions and the workings of the immune response (Cappuccio *et al.*, 2015). In addition, tracking the gene level response of *B. cinerea* can also identify any alteration in Arabidopsis gene levels directly caused by the pathogen.

This chapter also looked at the possibility of predicting key regulators of the Arabidosis defence response to *B. cinerea* from network structure. It would also be beneficial if, once the most accurate network of the stress has been created, it could be reduced to a size that makes simulating different conditions less computationally intensive and produces a small number of testable hypotheses. If a certain network characteristic exists that indicates a node is more likely to have an effect on the defence response, the network could be reduced to nodes exhibiting that characteristic. Manipulating those nodes *in vivo* would also be more likely produce the desired phenotype of increased disease resistance. However, no connection was found between the network indegree, outdegree, degree centrality, average shortest path length or hierarchical level of nodes and the probability of observing a phenotype when knocking out that node. This can be due to the fact that a full network consisting of genes and TFs has not been constructed. TFs which do not seem to be key in the TF-only network may in fact be hubs/bottlenecks in a network which contains all Arabidopsis genes.

While the same transcription factors may act in several different conditions, the interactions between regulators and targets in the network do change, as observed in yeast (Luscombe *et al.*, 2004). The overlap in transcriptome profiles between the response of plants to different abiotic stresses including heat, drought, cold, salt, high light or mechanical stress is also minimal (Mittler, 2006). The same can be said of the combined network for the 7 Arabidopsis stresses. This implies that while a TF can bind to many different targets, the genes it actually regulates depends on the specific circumstances. More than half of the hubs present in the common stress network were known stress-related genes. This supports the idea that TFs that are DE in multiple stresses, and have the same predicted edges in multiple stresses, are promising candidates in the search for TFs vital to the Arabidopsis stress response. Interestingly, a few of the top hubs of the individual stress networks were also identified in the other stresses as top hubs. These

were WRKYs and NF-YAs, which are well known regulators during biotic and abiotic stresses, as well as a BBX family TF, and a TIFY family TF. Not much is known about the latter 2 TFs but their identification as hubs suggests that further work should be done to explore their role in high light and *Pseudomonas hrpA* stresses (BBX14) and *B. cinerea* and drought stress (TIFY1).

In conclusion, GRN inference remains a complex problem in systems biology. While the network structure and a combined stress network did not reveal any major insights into the functioning of TFs, this may be due to the still intractable problem of inferring accurate networks. Incorporating data from multiple sources is preferable, as each method may a non-overlapping different set of regulatory links. Disparity between the networks suggested by the different data types can be caused to DNA methylation, chromatin accessibility, TF co-operativity and interactions of TFs with cofactors and transcription machinery (Franco-Zorrilla and Solano, 2017). However, incorporating information from experimental sources such as Y1H and ChIP-seq into models means that they cannot be used for validation purposes (in a fair trial, data for validation cannot be used to build the model). To solve this problem, the most robust network model had additional experimental data confirming direct regulatory links incorporated into it using the Borda count method. This model was subsequently used for modelling in Chapter 4.

2.5 Materials and Methods

2.5.1 *In silico* datasets

Time series data for *in silico* transcription factor networks is available from the Dialogue for Reverse Engineering Assessments and Methods 4 (DREAM4) *In Silico* Network Challenge archives at `https://www.synapse.org/#!Synapse:syn3049712/wiki/74628`. 5 networks of 10 and 100 nodes were used in this thesis: DREAM10.1 - DREAM10.5 and DREAM100.1 - DREAM100.5. The information provided in the DREAM4 archive includes time series training data, the gold standard network, mathematical descriptions of the Stochastic Differential Equations (SDEs) of the gold standard networks written in Systems Biology Markup Language (SBML), the nodes experiencing perturbation in the time series, and the strength of that perturbation. A full description of the methods used to generate the DREAM4 data is available in Marbach *et al.* (2009a).

2.5.2 Arabidopsis gene expression data

Seven time series datasets were available for the Arabidopsis transcriptome measured in different biotic and abiotic stress conditions: infection with *Pseudomonas syringae*, *Botrytis cinerea*, senescence, drought or high light. Each biological repeat consists of an Arabidopsis leaf exposed to the stress, from which the RNA was extracted for microarray hybridisation. The measurements were taken for each individual stress

or control condition using CATMA microarrays containing probes for the Arabidopsis transcriptome (Kurt *et al.*, 2005).

The dataset for infection with *Botrytis cinerea* (Windram *et al.*, 2012) contains 4 replicates of mock infected and infected measurements from CATMA arrays version 3 (30336 probes). Each replicate has gene expression measurements made every 2 hours over 48 hours after infection. There are two datasets for infection with *Pseudomonas syringae*: one with wild-type (WT) and one with a *hrpA*$^-$ mutant of *Pseudomonas syringae* (Lewis *et al.*, 2015) from CATMA arrays version 4 (32578 probes). The *P. syringae* dataset contains 4 replicates of measurements made in mock infection and infection with either strain of the bacterium. Each replicate contains 13 gene expression measurements taken over a period of 17.5 hours after infection.

The senescence datasets include a short day (unpublished data) and long day senescence studies (Breeze *et al.*, 2011) from CATMA version 3 arrays. There are 4 replicates for the short day senescence dataset, with measurements of gene expression taken every day for 19 days. There are 8 replicates for long day senescence, with measurements of gene expression taken every day for 11 days. The drought dataset (Bechtold *et al.*, 2016) was obtained using CATMA version 4 arrays. Daily measurements of 4 biological replicates are available for 13 days, for well-watered plants (95% soil water content) and plants in drought conditions (17% soil water content). The high light stress dataset (unpublished) consists of plants grown under a photosynthetic photon flux density of 1500 μmol m^{-2}s^{-1} from a light emitting diode and control plants grown in conditions of 150μmol m^{-2}s^{-1}. 4 replicates were sampled from each condition every 30 minutes for a total of 13 time points and hybridised to CATMA version 3 arrays.

From the available lists of differentially expressed genes (DEG) provided by the authors, only transcription factors were selected for use with network inference algorithms.

2.5.2.1 An Arabidopsis transcription factor list

An Arabidopsis transcription factor list was made by combining the lists reported by Pruneda-Paz *et al.* (2014) and an unpublished TF list generated by J. Moore using sequence similarity searches of DNA binding domains. TFs present in both lists were automatically included in the final selection, while those present in only one list were manually curated by inspecting the annotations of the THALEMINE (Krishnakumar *et al.*, 2016) and TAIR (Lamesch *et al.*, 2011) databases. Obsolete genes and those without transcription factor activity or the ability to bind DNA were removed from the list, giving a final list of 2,534 genes, or around 10% of the Arabidopsis coding genes. This list can be found in Dataset A.

2.5.3 Murine datasets

Murine gene expression time series data for 12 organs is available at the Gene Expression Omnibus (GEO) database, `www.ncbi.nlm.nih.gov/geo` (accession no. GSE54652) (Zhang *et al.*, 2014). Gene expression abundance was quantified using Affymetrix Mo-

Gene 1.0 ST arrays, every 2 hours for 48 hours. JTK-CYCLE (Hughes *et al.*, 2010) evaluation of all genes in all murine tissues was kindly forwarded to me by the authors. This consisted of 3186 genes with cycling transcripts. The time series data for these genes was used in the network inference analysis.

ChIP-seq data locating the binding sites for the BMAL1, CLOCK, NPAS2, PER1, PER2, CRY1, and CRY2 proteins in murine liver are available from Koike *et al.* (2012). These were used to create a gold standard network to determine the performance of the network inference algorithms using the murine time series.

A list of murine transcription factors was compiled from 3 databases: the Mouse Genome Database (Bult *et al.*, 2008), animalTFDB (Zhang *et al.*, 2011b) and the Riken Transcription Factor Database (Kanamori *et al.*, 2004). Redundant entries were removed, resulting in a list of 1403 transcription factors, given in Dataset B. There were used to inform the network inference algorithm as to which of the genes in the time series dataset were potential regulators. 356 of the 3186 genes with cycling transcripts were identified as transcription factors.

2.5.4 Experimental evidence for transcription factor-gene interaction

2.5.4.1 Unpublished datasets

A number of unpublished resources for TF-gene interactions were available from the BBSRC/EPSRC SABR (Systems Approaches to Biological Research)-funded Plant Responses to Environmental STress in Arabidopsis (PRESTA) project. Unpublished PRESTA data were kindly provided by Katherine Denby. This included Yeast-1-hybrid (Y1H) results and CATMA microarray (Kurt *et al.*, 2005) and NanoString (Kulkarni, 2011) data for knock-out (KO) and overexpressing (OE) Arabidopsis that were infected with *Botrytis cinerea*. As is common practice, Y1H results were used as a gold standard to rank different Arabidopsis network models (the disadvantages of doing this are explored in Section 2.4).

CATMA microarray data were available for 7 knock-out or overexpressing Arabidopsis lines and Arabidopsis Col-0 that had been infected with *Botrytis cinerea* (Table 2.16). Gene expression measurements were taken at one, two or three time points between 14 and 34 hours post infection (hpi). The expression data were filtered to select for genes that showed a 2-fold change in expression (either up-regulated or down-regulated) and a p-value$<$0.05. These DEG were assumed to be regulated (directly or indirectly) by the TF that was KO or OE. These data were used to rank the different Arabidopsis network models.

Background	AGI of KO gene
Col-0	-
anac092	AT5G39610
iaa17	AT1G04250
zat6	AT5G04340
rap2.6l	AT5G13330
anac055	AT3G15500
arf2	AT5G62000
wrky45	AT3G01970

Table 2.17: Backgrounds of the lines used in the PRESTA NanoString experiment. The gene AGI is the Arabidopsis Genome Initiative unique gene identifier, or locus code.

Gene	AGI	Mutant type	Time of measurement (hpi)
AT5G04340	*zat6*	KO	20, 26
AT3G06490	*myb108*	KO	26, 30, 34
AT3G06490	*myb108*	OE	30
AT2G22300	*camta3*	OE	26, 34
AT2G38470	*wrky33*	KO	14, 23.5, 25.5
AT5G37260	*cir1*	KO	26, 30
AT1G22070	*tga3*	KO	16, 24, 32
AT2G47190	*myb2*	KO	24, 30

Table 2.16: Transcriptomic datasets available from Arabidopsis mutants infected with *B. cinerea* from the PRESTA project. All gene expressions were measured with CATMA microarrays. The gene AGI is the Arabidopsis Genome Initiative unique gene identifier, or locus code.

Additionally, an additional PRESTA NanoString dataset consisted of gene measurements of 96 genes in different Arabidopsis backgrounds (see Table 2.17). The Arabidopsis leaves were infected with *B. cinerea* and the NanoString measurements were performed at 18, 22, 24, 26 and 32 hpi.

2.5.4.2 Published datasets

Yeast-1-hybrid and chromatin immunoprecipitation followed by sequencing (ChIP-seq) or DNA microarray analysis (ChIP-chip) data is publicly available at the Arabidopsis thaliana Regulatory network (AtRegNet) database (Yilmaz *et al.*, 2011) and the Arabidopsis Transcriptional Regulatory Map (ATRM) database (Jin *et al.*, 2015). AtRegNet contains regulatory links from published and unpublished data; the ATRM regulatory

links are obtained from on literature mining of transcriptional regulatory interactions. These were combined with the PRESTA Y1H data and used as a gold standard to rank different Arabidopsis network models.

2.5.5 Arabidopsis phenotypes

Observable Arabidopsis phenotypes during *Botrytis cinerea* infection following gene knock out or over-expression were collated from a number of sources. This ranged from published studies to unpublished PRESTA project data. A full list of the mutant genes, phenotype and source is given in Dataset E.

2.5.6 Network inference algorithms

The following network inference algorithms were used to obtain network models from DREAM network challenge data 2.5.1, Arabidopsis exposed to various biotic and abiotic stresses 2.5.2 and murine circadian genes 2.5.3. For each, all available time series data were used. Additionally all algorithms required a list of the genes which were TFs. In the case of the DREAM and Arabidopsis genes, all genes were TFs; only 356 of the 3186 murine genes were TFs. Further details on the mathematics underlying each algorithm is given in Appendix A. The networks inferred from the time series data of TFs from Arabidopsis leaves infected with *B. cinerea* are referred to as the *At-Botrytis* networks. Networks inferred for Arabidopsis exposed to various stresses are given as part of Dataset F.

2.5.6.1 Causal Structure Inference - CSI

The Causal Structure Inference network inference algorithm (Penfold and Wild, 2011) was run online on CyVerse UK (Polański *et al.*, 2017). The gene expression data were first linearly transformed so that the mean and standard deviation for each gene were 0 and 1, respectively. This was done using the z-score function in MATLAB, where the normalised data is transformed thus:

$$z = \frac{(x - \bar{X})}{S}, \tag{2.6}$$

where z is the normalised data, x is the vector containing gene expression data, $\bar{X}$ is the mean of the data and S the standard deviation.

The normalised data, together with the list of TFs in the dataset, was then submitted to the CSI algorithm on CyVerse UK. The default running parameters were used: parental set depth of 2, a gamma prior on the Gaussian process (GP) hyperparameter values ($\Gamma(10; 0.1)$) and a weight truncation of 10^{-5}. The resulting matrix was a fully connected marginal network model, with a probability score for each regulator-target interaction.

2.5.6.2 GENIE3

The MATLAB version of the GENIE3 network inference algorithm (Huynh-Thu *et al.*, 2010) was downloaded from `http://www.montefiore.ulg.ac.be/~huynh-thu/GENIE3.html`. The Random Forest method was run on each time series dataset to generate an ensemble of 1000 trees with K attributes being selected randomly to determine the best split, where K is the square root of the number of input transcription factors. These particular parameters were selected because they were best at predicting the *E. coli* network model in (Marbach *et al.*, 2012).

2.5.6.3 GRENITS

The GRENITS (Morrissey, 2013) package for R was downloaded from `https://bioconductor.org/packages/release/bioc/html/GRENITS.html`. The inputs used were the time series data and list of TFs. A dynamic Bayesian network of the linear interactions between regulators and targets is generated by first running the Markov chain Monte Carlo (MCMC) function with default parameters to obtain a posterior distribution for the parameters of the Bayesian model. The network probabilities are then obtained by running the *analyse.output* function.

2.5.6.4 Inferelator

The code for running the Inferelator network inference algorithm (Madar *et al.*, 2010) was kindly provided by Daniel Marbach. The pipeline involved running time-lagged Context Likelihood of Relatedness initially and having the resulting network refined by Inferelator. The Inferelator R code was modified slightly and is provided in Dataset C. No cut-off for number of edges was specified as the thresholding was applied later. The number of bootstraps for estimating regulatory interactions was 100 and the maximum number of regulators considered in the elastic net step was 30.

2.5.6.5 TIGRESS

The code for running the TIGRESS network inference algorithm (Haury *et al.*, 2012) was kindly given by Daniel Marbach. Parameter values used were R=500; alpha=0.3; L=3 for the number of stability selection runs, the perturbation parameter and the number of LARS steps, respectively. The purpose of each parameter is further explained in Appendix A. The original stability scores were used to calculate an area score ranking the importance of a TF based on their influence on the target gene.

2.5.7 Network consensus

A consensus network was calculated using different models as input with the AverageRank online tool from Gene Pattern `http://dream.broadinstitute.org/gp/pages/index.jsf`, which is based on the Borda count method, and is introduced in (Marbach *et al.*, 2012). In order to incorporate biological evidence as prior for a network

model, a modification on the Borda count method was used to give additional weight to interactions that were experimentally validated using methods such as Y1H and ChIP-seq (but not KO or OE data).

The top 100,000 interactions from the *At-Botrytis model* were first scored so that the interaction with the least confidence received 1 point, the interaction just above that received 2 points and so on, with the most top ranked interaction receiving 100,000 points. Interactions that had prior evidence from experiments received an additional 5,000 points. The interactions were then sorted based on this point system, with the top ranked interaction receiving the most points, and so on.

2.5.8 Network ranking using AUROC and AUPR

The area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR) were calculated for the DREAM network models using a MATLAB function by Stefan Schroedl, 2010. AUROC is the area under the true positive ratio vs false positive ratio curve and AUPR is the area under precision vs true positive ratio curve. The inputs used for this function were an adjacency matrix for the gold standard and a weighted adjacency matrix generated from one of the network inference algorithms.

2.5.9 Bayesian Information Criterion

In order to calculate the Bayesian Information Criterion for a certain network structure, a Gaussian process was fit to the time series dataset, given that particular network model. A different model meant that a gene could be regulated by a different set of regulators. Hyperparameters were optimised for the squared exponential function as explained in Section 4.6.3.1. The negative log likelihood was then returned for these optimised parameter values using the *gp* function from GPML. The *aicbic* function in MATLAB was used to calculate the BIC score given the maximal likelihood for the model and the network model size (number of edges).

2.5.10 Arabidopsis network ranking using biological data

In order to determine which of the TF network models for Arabidopsis infected with *Botrytis cinerea* were the most accurate and/or precise, biological data obtained from yeast-1-hybrid (Y1H), chromatin immunoprecipitation followed by sequencing (ChIP-seq) or microarray (ChIP-chip), electrophoretic mobility shift assays, and gene expression data from knock-out and overexpressing plants were used to rank the CSI, GENIE3, GRENITS, Inferelator, and TIGRESS-produced models.

2.5.10.1 Edge verification with Y1H and ChIP data

The *At-Botrytis* network models were first obtained using one of the network inference algorithms in Section 2.5.6. A threshold was applied to each model so that interactions with a score below that particular threshold were eliminated from the model.

The threshold was varied from 0.00001 to 0.001 in increments of 0.00001, 0.001 to 0.1 in increments of 0.001, and 0.1 to 1 in increments of 0.01. This range of thresholds was chosen in order to encapsulate models with 0 edges (a stringent threshold, ∼0.1 to ∼1) all the way up to $> 100,000$ edges (a lenient threshold ∼0.00001 to ∼0.001). The number of edges at a particular threshold is referred to as the network size.

Biological evidence for the direct interactions between TFs and promoters were obtained from several source as described in Section 2.5.4: Y1H, ChIP-seq and ChIP-chip from the PRESTA project, and the ATRM and AtRegNet databases. This was used to construct an adjacency matrix of experimentally-confirmed interactions for the 883 TFs in the *At-Botrytis* network (provided in Dataset D).

The network models obtained for each algorithm for each threshold were then compared to this adjacency matrix. An interaction between a TF, X, and the promoter of a TF, Y in the network models counted as a true positive if this interaction was also observed in biological experiments. The total number of true positives was calculated for each network size.

2.5.10.2 Edge verification using Arabidopsis knock-out/over-expression transcriptomic data

Transcriptome data from CATMA array measurements were available for 7 single gene knock-out (KO) and over-expression (OE) experiments from the PRESTA project. The plants were all infected with *B. cinerea* during the transcriptome measurements. Wild-type plants infected with *B. cinerea* were used as the control. These datasets contained log fold change values and adjusted p-values for genes in the knock-out or over-expression line compared to the WT line. The microarray measurements were available for either single time points, or 2/3 time points after infection with *B. cinerea*. The genes with a log fold change of more than 1, and a p-value of less than 0.01, were selected as differentially expressed between the mutant and WT. Where several time points for the same knock-out/over-expressor were available, all the differentially expressed genes from all the time points were included. In addition a list of WRKY33-influenced TFs was available from Liu *et al.* (2015) at 14 hpi, which have also been taken into account.

NanoString gene expression measurements for 96 genes from 3 KO lines were also used: *rap2.6l, anac055, arf2*. These measurements were all made at 18, 22, 24, 26, 32 hpi in plants infected with *B. cinerea*. Differentially expressed genes were determined for the NanoString data by comparing the mean expression of each gene at a time point in the KO experiment +/- 2 standard deviations with the expression of the same gene at the same time point in the WT control experiment +/- 2 standard deviations. If these two ranges did not overlap, the gene was considered to be differentially expressed. The DEG in all the different lines are available as Dataset D.

The differentially expressed genes in the KO/OE experiments were limited to those which are also present in the *At-Botrytis* network. For a knock-out/over-expression experiment of transcription factor X, the differentially expressed genes must either be directly regulated by TF X, or indirectly regulated, through intermediary transcription factors. Therefore a true positive in an *At-Botrytis* network model was counted if a DEG

in the KO/OE experiment of TF X was also present downstream (maximum of three degrees downstream) of that TF X.

The number of true positives was counted for each network model at a range of network sizes, similar to Section 2.5.10.1. MATLAB code for this purpose is provided in Dataset G.

2.5.11 Network visualisation

Cytoscape (Shannon *et al.*, 2003) was used to visualise networks, using an input list of regulators and their targets. The Cytoscape application, Pesca (Scardoni *et al.*, 2015), was used to find all the shortest paths between two nodes in a network.

2.5.12 Network and node characteristics

Network adjacency matrices were used of size n x n, where n is the total number of nodes in the network. Regulators are found in columns and targets in rows - an entry in cell (2,3) of the matrix means that gene 3 regulates gene 2. Entries in matrices consist either of edge scores as obtained from the network inference methods, which can be converted to Boolean once a threshold has been applied - scores above a threshold x become 1 and those below the threshold become 0.

The outdegree of Gene X in a network was determined by summing the rows in column X of the Boolean adjacency matrix. The indegree of Gene X was determined by summing the columns in row X of the Boolean adjacency matrix. The node degree was obtained by adding the indegree and outdegree of the node. The clustering coefficient is calculated in MATLAB using code by David Gleich, 2008. The indegree, outdegree, betweeness and pagerank centrality of nodes were determined using the MATLAB centrality function.

Histograms were created with the histogram function in MATLAB. Correlation plots were created using the MATLAB function corrplot. Hypergeometric tests were carried out using the MATLAB function hygecdf.

2.5.12.1 Characterising nodes that produce phenotypes

MATLAB code for producing network hierarchies is provided in Dataset G.

2.5.13 Network hierarchy construction

Breadth-first search was employed to create a hierarchy of the nodes in the *At-Botrytis* network as per Yu and Gerstein (2006). Additionally, a second hierarchy was produced as per Bhardwaj *et al.* (2010). The steps were the same as in Yu and Gerstein (2006), with an extra step to prevent loops. When TFs were assigned to levels 1 and above, any unassigned TFs that they regulated were also assigned to their level.

The number of genes which give rise to a phenotype (increased or decreased resistance to *B. cinerea* infection) upon perturbation and the number of genes which do not give rise to a phenotype when perturbed are tallied at each hierarchical level. A

hypergeometric test was used to determine whether there was an unusual enrichment for positive or negative regulators of defence at any hierarchical level.

Chapter 3

A Framework for Engineering Stress Resilient Plants using Genetic Feedback Control and Regulatory Network Rewiring

3.1 Aims

- Identify a subnetwork of the *At-Botrytis* transcription factor (TF) gene regulatory network (GRN).
- Identify genes within this subnetwork, the 9GRN, which play a role in the Arabidopsis defence to *B. cinerea*.
- Parameterise an ordinary differential equation (ODE) model for the 9GRN.
- Implement an idealised feedback controller to counteract the effects of perturbation to a positive regulator of defence, CHE, based on work by Harris *et al.* (2015).
- Design the feedback controller solely through the manipulation of edges in the 9GRN to restore the levels of *CHE*.

3.2 Acknowledgements

This chapter is based on the manuscript Foo *et al.* (2017), authored by Mathias Foo, Peijun Zhang, Declan Bates, Katherine Denby, and myself. Yeast-1-Hybrid data and *at-ERF1* lines were generated by PZ. MF designed and implemented the feedback controller for the 9GRN model, and performed the robustness analysis. I created the network models, collected biological data for validation of the model, inferred a dynamic network model using ODEs, and discussed the implementation of such a circuit *in vivo*. MF and I constructed the rewired network models.

3.3 Introduction

The network models inferred in the previous chapter attempt to summarise the complex regulatory interactions of the Arabidopsis TF defence system. These top-down approaches offer a great overview of the system; however they provide little detail on the richness of the molecular interactions. Ultimately, the aim is to predict the behaviour of the system under novel rewiring conditions. For this, coupled ODEs are used to represent smaller genetic circuits from the large network and to optimise gene expression of its constituent parts. ODEs are a modelling tool borrowed from engineering; the use of control to enforce desired behaviour in a system is also a large part of engineering and its principles are applied in this chapter to the Arabidopsis GRN system. This allows the optimisation of expression levels of genes in the GRN undergoing perturbation.

Proportional integral (PI) controllers are a common tool in engineering for minimising the error between a desired system output (the setpoint) and the actual output. These consist of two basic types of control: proportional control, which considers the current error and integral control, which considers the sum of past errors. Through the action of the integral component, the system is guaranteed to reach the desired output without a steady-state error. Mathematically, the PI controller can be described thus:

$$u(t) = K_p e(t) + K_i \int_0^t e(t)dt \tag{3.1}$$

where u is the control signal, e is the error, and the controller parameters K_p for proportional gain and K_i for integral gain.

A general PI controller takes a form shown in Figure 3.1.

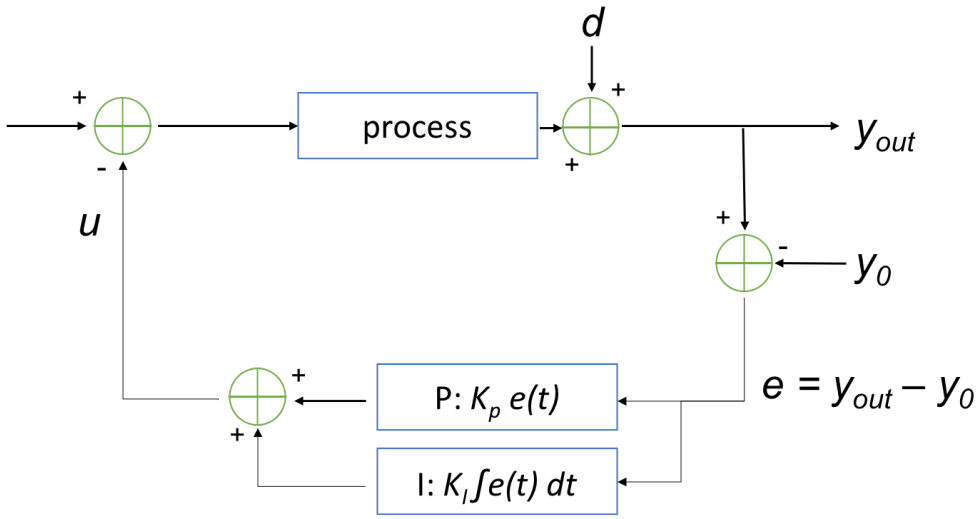


Figure 3.1: Circuit diagram of a typical PI-controlled system with the input signal and the output signal y_{out}. The error signal, e is the difference between the output signal and the setpoint, y_0. This difference is fed back through the PI controller which then produces some control action signal, u, that is subtracted from the input. The PI controller can reject the effects of d, a disturbance to the system. The controller modules, P and I have parameters K_p and K_I for proportional and integral gain, respectively. These parameters can be tuned to optimise the performance of the system.

Integral control has also evolved in biological systems. Yi *et al.* (2000) showed that integral feedback control is necessary for biological systems to achieve perfect adaptation, a condition where the output is independent of the input during steady state. This can be seen in *E. coli* chemotaxis. Receptors on the surface of the cells detect molecules, such as sugar, and adjust the ratio of running and tumbling movements by changing their flagellar motion. By decreasing the likelihood of tumbling, the cells are able to run in a straight line up a concentration gradient towards the chemoattractant. Prolonged stimulation of receptors leads to their methylation, restoring the balance between the running and tumbling ratios. Barkai and Leibler (1997) proved that perfect adaptation is a characteristic of the system design rather than the parameters. In line with this deduction, this chapter introduces adaptation to a system facing perturbation by constructing a specific network structure through rewiring to provide control.

Unfavourable environmental conditions during the growth of crop plants can cause significant yield loss and reduction in quality. These conditions include abiotic stresses, such as drought and extreme temperature, as well as the biotic stresses of disease and herbivory. Climate change is driving increasingly unpredictable and variable weather, and bringing associated change in pathogen (and hence disease) prevalence and incidence (Bebber *et al.*, 2013, 2014). It is therefore important to develop crops that are

resilient to varying conditions and able to maintain yield in suboptimal environments (Buchanan-Wollaston *et al.*, 2017). The introduction and/or removal of single genes via genetic engineering has led to plants with enhanced tolerance to particular abiotic and biotic stresses (Parmar *et al.*, 2017), however, often such approaches have unintended consequences on other plant responses, (Veronese *et al.*, 2006) and in the case of disease resistance they may not be durable.

Recent increased understanding of how plant responses to different environmental conditions are controlled and integrated, together with the development of systems biology approaches, has opened up the possibility of designing stress resilient crops using engineering principles. In this work, we have focused on transcriptional regulation, as transcriptional reprogramming is a significant component of plant stress responses (Wilkins *et al.* (2016), Lewis *et al.* (2015), Breeze *et al.* (2011)) and a point of cross-talk between responses to different stresses (Sharma *et al.*, 2013).

Here, we focus on the regulation of the defence response induced in Arabidopsis by the fungal pathogen, *Botrytis cinerea* (Windram *et al.*, 2012). When pathogens infect plants, disease is the result of dynamic interactions between the two organisms. Pathogens secrete a range of proteins, small RNAs and metabolites to disrupt host defence and manipulate the extra- and intracellular environment to aid colonisation (Williamson *et al.* (2007), Jamir *et al.* (2004), Weiberg *et al.* (2013), Jones and Dangl (2006)). This is thought to explain why some positive regulators of defence are down-regulated during infection, for example expression of *TGA3* decreases during *B. cinerea* infection of Arabidopsis, yet plants lacking *TGA3* expression are more susceptible to this pathogen (Windram *et al.*, 2012). In this study, we use a control engineering approach to counteract such potentially pathogen-mediated perturbations of positive regulators of defence. Constitutive overexpression of such positive regulators would be an obvious approach, but this brings significant drawbacks; the positive regulator of defence may have other roles in the plant which are disrupted due to constitutively high levels of expression, and constitutive activation of plant defence responses is known to often impact on growth (Heidel *et al.*, 2004). Our proposed approach, which seeks to dynamically respond to perturbations of expression over the time-course of infection, should overcome these drawbacks.

From the perspective of control engineering, this scenario can be naturally formulated mathematically as a disturbance attenuation problem. For such problems, control engineers have developed a variety of powerful theoretical tools and techniques that allow the design of feedback controllers that can attenuate the effects of external perturbations on the functioning of a system or network (see Hespanha *et al.* (2007) and references therein). The application of these tools to the analysis and design of complex biological networks is now attracting significant interest within the synthetic biology community (Liu *et al.*, 2011; Vinayagam *et al.*, 2016). To date, however, the potential usefulness of such approaches for engineering more resilient plants has not been investigated.

Here, we explore how combining control engineering design tools (Ang *et al.* (2010), Briat *et al.* (2016b), Harris *et al.* (2015)) with synthetic biology techniques could be used to enhance resistance against *B. cinerea* in Arabidopsis by preventing down-

regulation of a positive regulator of defence during infection. We design and test our controller using a model of the Arabidopsis gene regulatory sub-network underlying the transcriptional response to *B. cinerea* infection. This network model is formulated using ordinary differential equations (ODEs) and constructed from experimental data using network inference and system identification techniques. It is then validated against different time-series transcriptome datasets capturing the response of the plant's regulatory network to pathogen attack. Simulation results show the capability of the proposed approach to significantly reduce the perturbation of a positive regulator of plant defence in response to infection. We propose a novel strategy for implementing the controller experimentally, which avoids the need for the incorporation of any exogenous synthetic control circuitry. This strategy is based on the insight that the network motif required for the controller can be implemented by rewiring the regulatory regions of existing genes in the plant's stress-response network. We show how this can be done through the addition of gene coding sequences under the control of alternative regulatory regions.

3.4 Results

3.4.1 Inferring the regulatory sub-network containing a positive regulator of defence.

We previously generated a high-resolution time series of the Arabidopsis transcriptome during the first 48 hours after inoculation by the pathogen *B. cinerea* (Windram *et al.*, 2012). Nearly 10,000 genes were identified as being differentially expressed in infected leaves compared to mock-inoculated leaves, including 883 TFs. We used the time-series transcriptome data for the differentially expressed TFs as input for network inference algorithms, to generate causal directed network models of the regulatory events underlying changes in expression of these TF genes. The algorithms chosen for this purpose (GENIE3 (Huynh-Thu *et al.*, 2010), TIGRESS (Haury *et al.*, 2012) and Inferelator (Bonneau *et al.*, 2006)), were highly ranked in a recent assessment of network inference algorithms (Schaffter *et al.*, 2011). GENIE3 approaches network inference as a tree-based regression problem and came first in the DREAM4 *in silico* multifactorial network inference challenge (Schaffter *et al.*, 2011). Inferelator and TIGRESS both use feature selection and least angle regression to rank the potential regulators of a gene. The outputs from these three algorithms were used to generate a consensus network model, as a robust way of generating high confidence networks (Marbach *et al.*, 2012). A threshold (edges $\leq$ 10 times the number of nodes) was applied to this consensus network to limit it to 8,830 edges. Furthermore only the top three regulators of each node were kept based on the highest probability score. From this final network, we looked for sub-networks surrounding positive regulators of defence against *B. cinerea* that were downregulated during infection. Genes that are positive regulators of defence play a positive role in the response to stress, and are identified from experiments if they produce more susceptible plants when knocked out, or more resistant plants when over-expressed. Negative regulators of defence create more resistant plants when knocked out, or more susceptible

plants when overexpressed. This led us to focus on a 9-gene regulatory network, termed 9GRN (see Figure 3.4) containing the TF *CCA1 HIKING EXPEDITION* (*CHE*), which includes predicted upstream regulators of *CHE*. The 9GRN was a sub-network whereby all nodes were highly connected to each other, but not to the rest of the network, and which contained two genes which led to a visible stress phenotype when knocked out. (Caveat - this was carried out before of some the work indicating in Chapter 2 that the overall consensus model is the best to use, rather than the consensus between GENIE3, Inferelator and TIGRESS).

3.4.2 *CHE* is a positive regulator of defence and *at-ERF1* is a negative regulator of defence against *B. cinerea*.

Expression of the transcription factor (TF) *CCA1* HIKING EXPEDITION (*CHE*) is downregulated during *B. cinerea* infection (Windram *et al.* (2012) and Figure 3.2a). Rhythmic expression of *CHE* is clear in the mock-inoculated samples (reflecting the role of *CHE* within the circadian clock (Pruneda-Paz *et al.*, 2009)) with downregulation due to infection beginning around 22 hours post inoculation. A mutant with significantly reduced expression of *CHE*, *che* − 1, (Pruneda-Paz *et al.*, 2009) shows increased susceptibility to *B. cinerea* compared to wildtype indicating *CHE* plays a positive role in defence against this pathogen (Figure 3.2b). In addition to *CHE*, two other genes in the 9GRN are important in defence against *B. cinerea*: *ORA59* and *at-ERF1*. *ORA59* is a positive regulator of defence (Pré *et al.*, 2008), while *at-ERF1* is a negative regulator of defence (Figure 3.3). The other genes in the network include *MYB51*, which regulates the synthesis of the secondary metabolite indolic glucosinolate (Gigolashvili *et al.*, 2007), *LOL1*, a positive regulator of cell death through reactive oxygen species (Epple *et al.*, 2003), *ANAC055*, a regulator of the jasmonic acid defence response and abscisic acid signalling (Jiang *et al.*, 2009), *ATML1*, a homeobox gene controlling differentiation in Arabidopsis embryo development (Takada *et al.*, 2013), *RAP2.6L*, which responds to multiple stress hormones such as salicylic acid, jasmonic acid, abscisic acid, ethylene, as well as participating in signalling during salt and drought stress (Krishnaswamy *et al.*, 2011), and *AT1G79150*, about which not much is known.

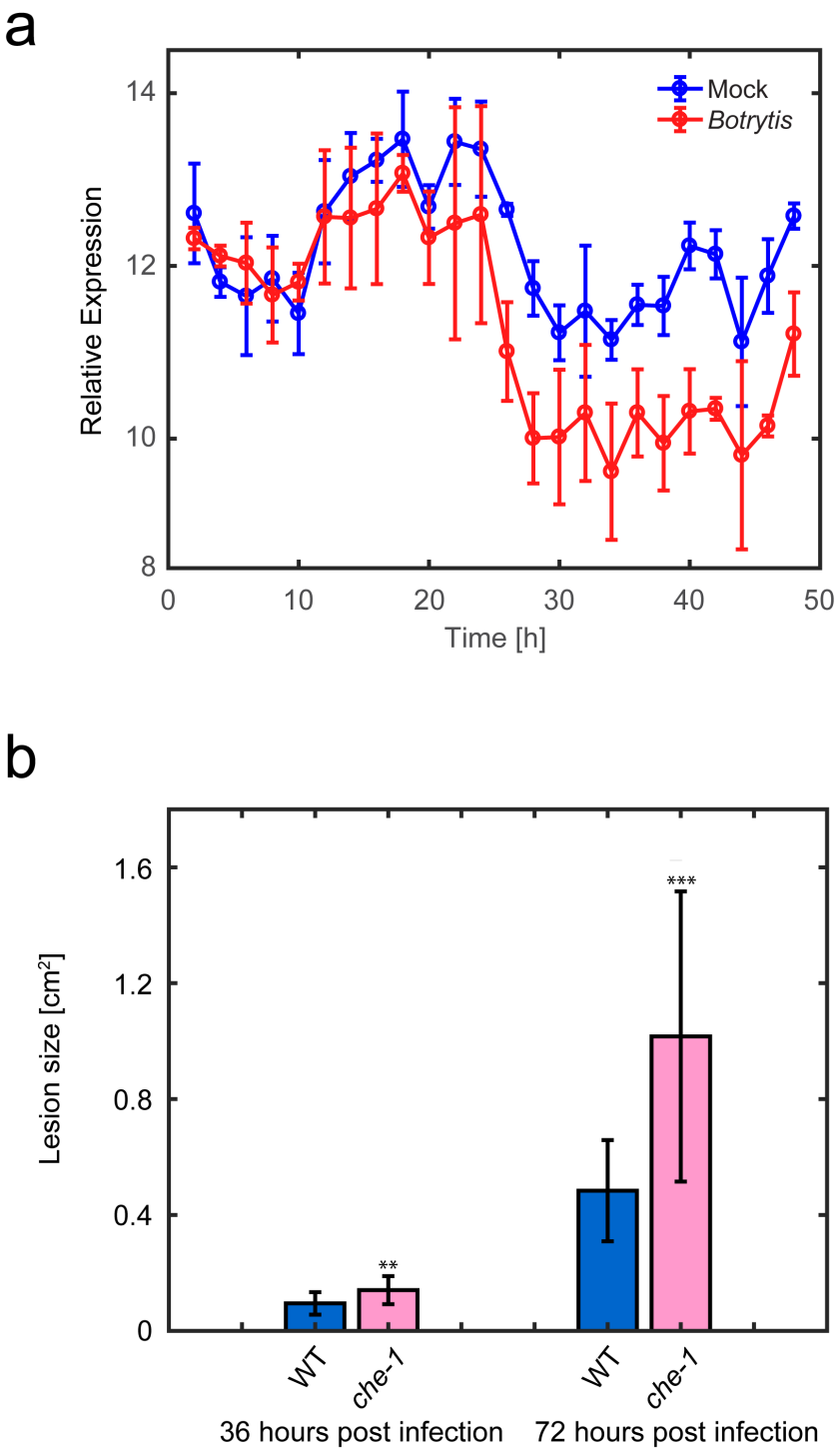


Figure 3.2: Expression and role of *CHE* during infection with *B. cinerea* (a) Expression of the TF *CHE* is downregulated during *B. cinerea* infection of Arabidopsis leaves. Leaves were drop-inoculated with *B. cinerea* spores or mock-inoculated, and genome-wide gene expression determined every 2 hours for both mock treatment (blue) and *B. cinerea* infection (red).

Figure 3.2: *(previous page)*: Open circles are the average of four biological repeats with bars representing standard deviation. This data is extracted from Windram et al, 2012. (b) CHE is a positive regulator of defence against *B. cinerea*. Lesion size of Arabidopsis leaves (n = 17) drop-inoculated with fungal spores were measured 36 and 72 hours post infection. *che-1* is an Arabidopsis mutant with significantly reduced *CHE* expression. WT is the wildtype Col-0 Arabidopsis accession. Error bars represent standard deviation, ** represents $p \leq 0.01$ and *** represents $p \leq 0.001$.

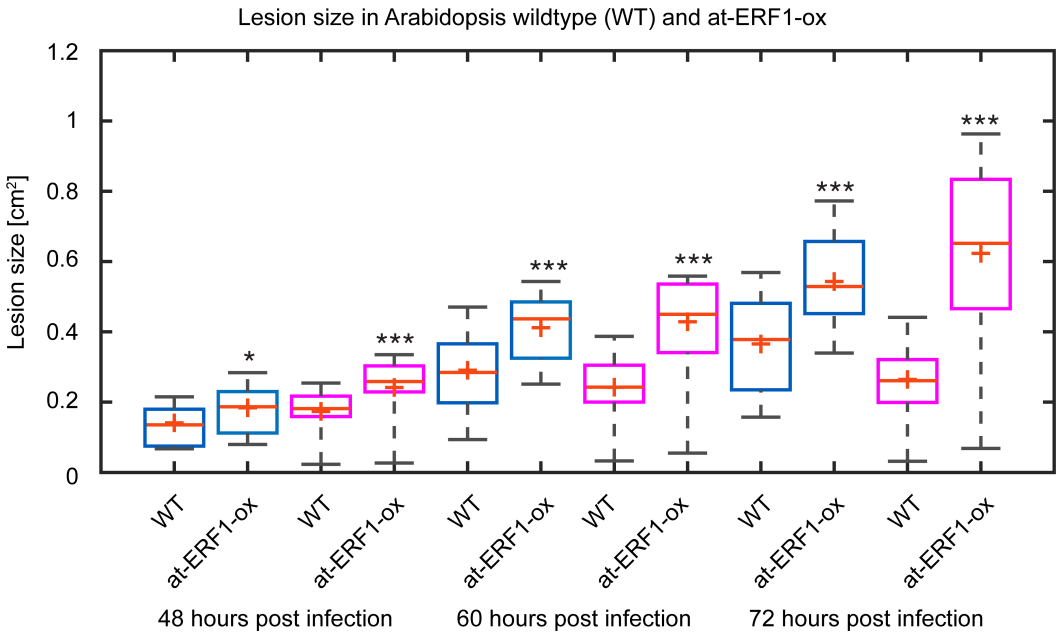


Figure 3.3: at-ERF1 is a negative regulator of defence against *B. cinerea*. Box and whisker plots show the lesion size in wildtype Arabidopsis leaves and leaves from an Arabidopsis mutant overexpressing *at-ERF1*, after inoculation with *B. cinerea*. Each box and whisker plot represents measurements from 30 leaves at 48, 60 or 72 hours after the start of infection. Crosses represent the mean lesion size. * represents a p-value ≤ 0.05 and *** represents $p \leq 0.001$. Blue and pink box outlines denote separate independent experiments.

3.4.3 Validating edges in the 9GRN model.

To increase our confidence in the validity of the inferred 9GRN sub-network model, we used yeast-1-hybrid (Y1H), a partial Arabidopsis cistrome map (O'Malley *et al.*, 2016), and gene expression data from *RAP2.6L* overexpressors (Hickman *et al.*, 2017) to test regulation predicted by the model. A set of pair-wise Y1H had been carried out testing binding of 75 TFs to the promoter regions of 34 of the same TFs within the PRESTA project. Within this set, there were 4 edges in our model (*RAP2.6L*

to *ANAC055*; *ANAC055* to *RAP2.6L*, *ANAC055* to *ORA59* and *at-ERF1* to *ORA59*) that had been tested. For two of these edges, strong binding was seen in the Y1H experiments; *RAP2.6L* could bind to the promoter of *ANAC055* and *at-ERF1* could bind to the promoter of *ORA59* (Figure 3.5). In addition, the Y1H data suggested two additional edges that were missing from our model (*RAP2.6L* to *ORA59*, and *ORA59* to *ANAC055*), however, expression data from *RAP2.6L* overexpressors (Hickman *et al.*, 2017) and knockout mutant of *ORA59* (Pré *et al.*, 2008) do not show any evidence for these regulatory edges. Additional interactions in the 9GRN were verified using data from an Arabidopsis cistrome map (O'Malley *et al.*, 2016). The cistrome is the complete set of cis-elements or TF binding sites in an organism, and a partial map was generated by O'Malley *et al.* (2016) using DNA affinity purification sequencing (DAP-seq) to identify TF binding sites for 349 TFs (including *CHE*, *ORA59*, *ANAC055* and *MYB51* from our network). This analysis revealed that *ANAC055* can bind to the promoters of *ORA59* and *RAP2.6L*. Finally, the *RAP2.6L* overexpressing mutant showed increased expression of *AT1G79150*, providing evidence for this regulatory interaction (Hickman *et al.*, 2017). Edges with supporting experimental data are shown in green in Figure 3.4.

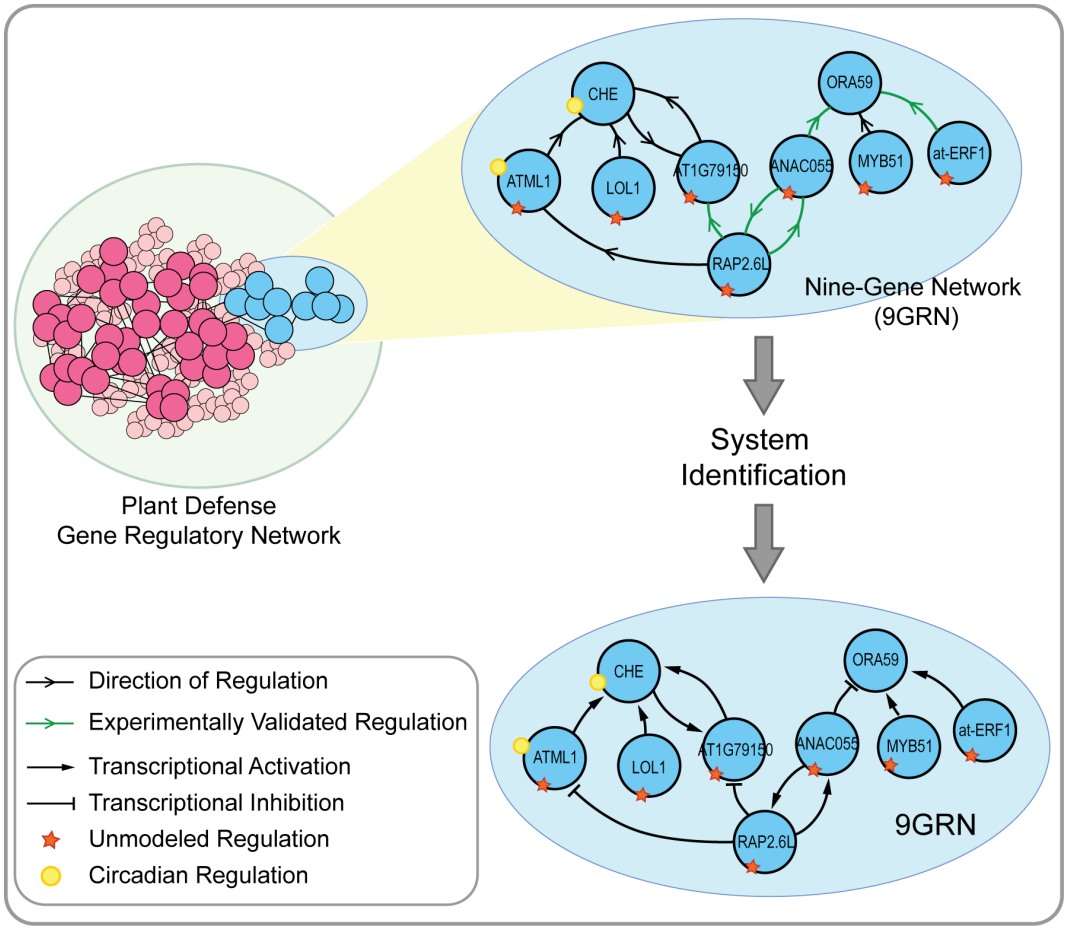


Figure 3.4: Network model of gene regulatory events mediating transcriptional response to *B. cinerea*. The nine-gene network (9GRN) is a sub-network of the initial network model inferred from time series transcriptome data. The direction of regulation is indicated by the arrow. Red stars represent unmodeled regulation (e.g. direct regulation from *B. cinerea*, noise and other unidentified regulation). The yellow circle represents circadian regulation. Green edges represent interactions that are supported by experimental data. The regulation types (arrow-head and bar-head) in 9GRN are identified through system identification.

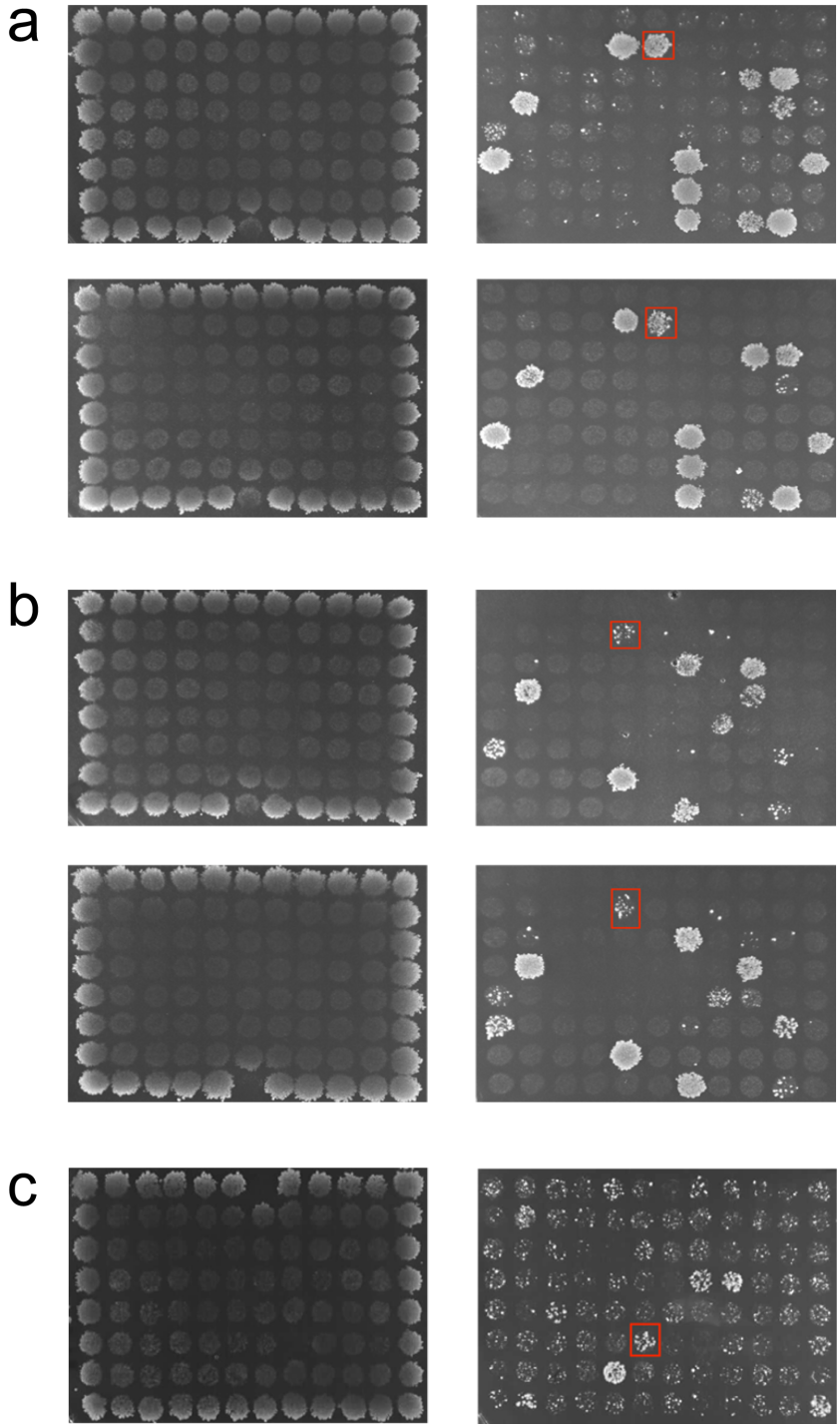


Figure 3.5: Yeast-1-hybrid results for at-ERF1 binding to *ORA59* (a), RAP2.6L binding to *ANAC055* (b), ANAC055 binding to *ORA59* (c). The first column shows growth of diploid yeast on the positive control unselective media and the second column shows growth on selective media. The relevant interactions are highlighted in red. Different rows are experiments carried out on different days

3.4.4 A validated dynamic model of the CHE regulatory sub-network.

The network inference algorithms used to infer the large consensus network model are able to predict regulatory relationships between the genes in the 9GRN but the type of regulation (i.e. activating or inhibiting) cannot be determined. Since these are essential features of any model that can be used for controller design, we next determined the direction of the regulatory edges in the 9GRN using standard four-step system identification techniques: data collection, model structure selection, parameter estimation and model validation (see chapters 1 and 7 of Gardner *et al.* (2003)). Previous studies that utilised this technique to identify regulation types in GRNs used linear models (Gardner *et al.* (2003), di Bernardo *et al.* (2005), Bansal *et al.* (2007)), and there is now strong evidence that the underlying dynamics of GRNs can be accurately described using such models (we define accurate as a model able to recapitulate experimental data within a single standard deviation of error) (Dalchau *et al.*, 2011; Herrero *et al.*, 2012). Non-linear models are able to capture more behaviours compared to linear models, at the cost of computational time and parsimony. Our previous work on simulating rewiring an *in vitro* DREAM4 challenge network (Schaffter *et al.*, 2011) showed a linear model was sufficient to capture the dynamics of the system and was not improved upon by a non-linear S-System model (Foo *et al.*, 2017). Moreover, as the model is subsequently to be used to design perturbation mitigation strategies, a linear model facilitates the use of linear control design techniques that are more established than their nonlinear counterparts. In system identification terminology, black box models refer to a set of ready-made models with no physical structure or biological interpretation. On the other hand, gray box models refer to models that are tailor-made given some prior information about the system. Since we have prior knowledge of the direction of regulation between the genes obtained from the inferred network above, we use a linear gray box model comprising nine ODEs (Equation in Section 3.6.6) for the 9GRN, and thus only need to identify the regulation type and dynamics within the 9GRN.

The values of the model parameters were estimated from the available mRNA time-series data (Windram *et al.*, 2012) using a nonlinear least squares algorithm (Kim *et al.* (2007), Kim *et al.* (2008b)) (Equation 3.8 in Methods section) and the estimated parameters are given in Table 3.3 in Methods section. As these mRNA time-series measurements are normalised using an intensity-dependent normalization method, (Wu *et al.* (2003), Yang *et al.* (2002)) the resulting measurements are dimensionless and are reported as relative expression. Figure 3.4 indicates the regulation types identified in the 9GRN sub-network, where positive and negative values of production rate given in Table 3.3 in the Methods section denote transcriptional activators and inhibitors respectively. In addition, all the estimated degradation rates had the expected negative sign, and had numerical values within the range expected (Narsai *et al.*, 2007). We validated the dynamic model by comparing its response against another mRNA time-series data set (see Methods section) that was not used in the parameter estimation process as well as two mutant behaviors. As shown in Figure 3.6, the identified model is able to accurately predict the expression behavior of the network. Additionally, the model shows good predictive capability against two mutant datasets (see Figure 3.7). These two datasets

were not used for parameter estimation and therefore emphasise the good predictive capability of the simple linear model.

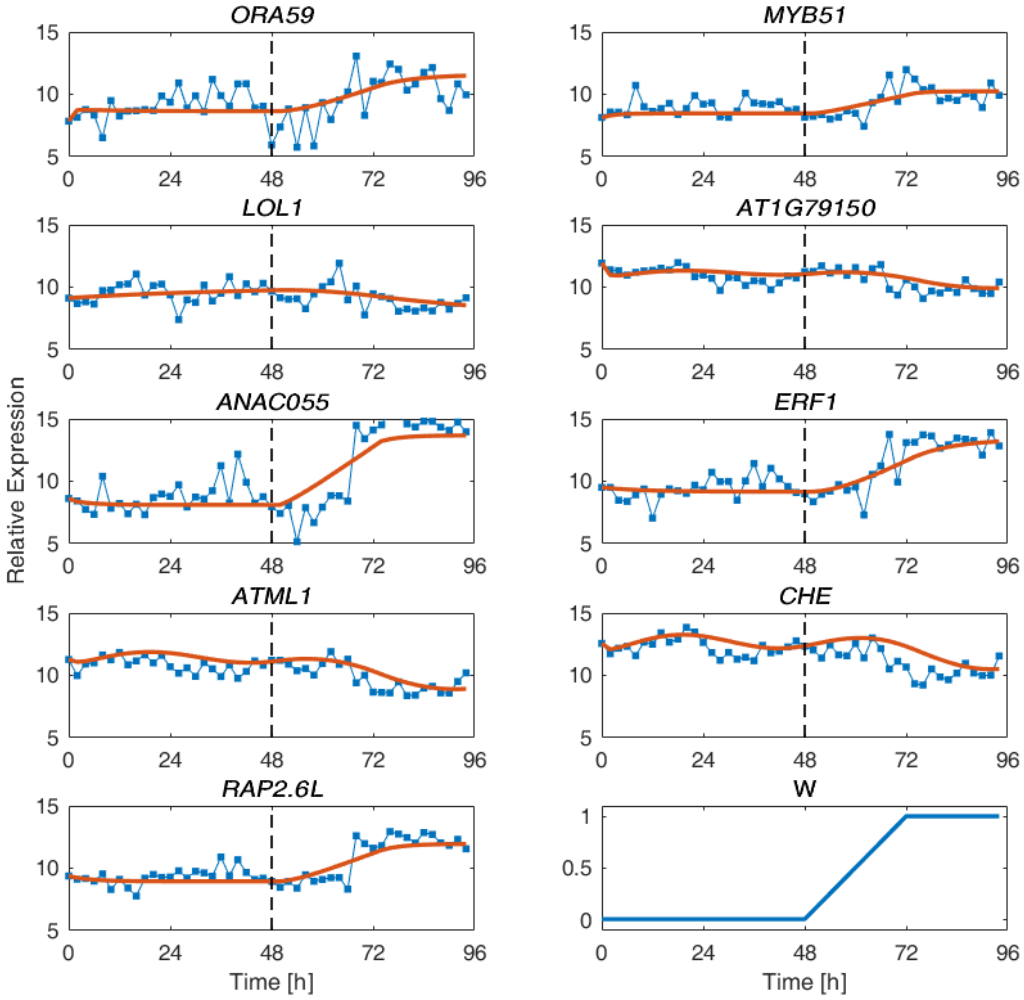


Figure 3.6: Validation of the linear model against an experimental data set that was not used in the parameter estimation exercise. The experimental data sets in Windram *et al.* (2012) are composed of two time series, one mock-inoculated and one *B. cinerea*-inoculated. Here, these two time series are joined (denoted by the vertical dashed line) to illustrate a transition from pre- to post-infection, with *B. cinerea* infection starting at time 48 hours. There are four sets of such joined time-series data; we used the average of the first three data sets for parameter estimation, leaving the fourth data set for model validation shown above. We have also included the unmodeled regulation, W (perturbation due to *B. cinerea* infection, noise and any other regulation) described by Equation 3.6. Line with dots: Experiment data, Solid line: Linear model.

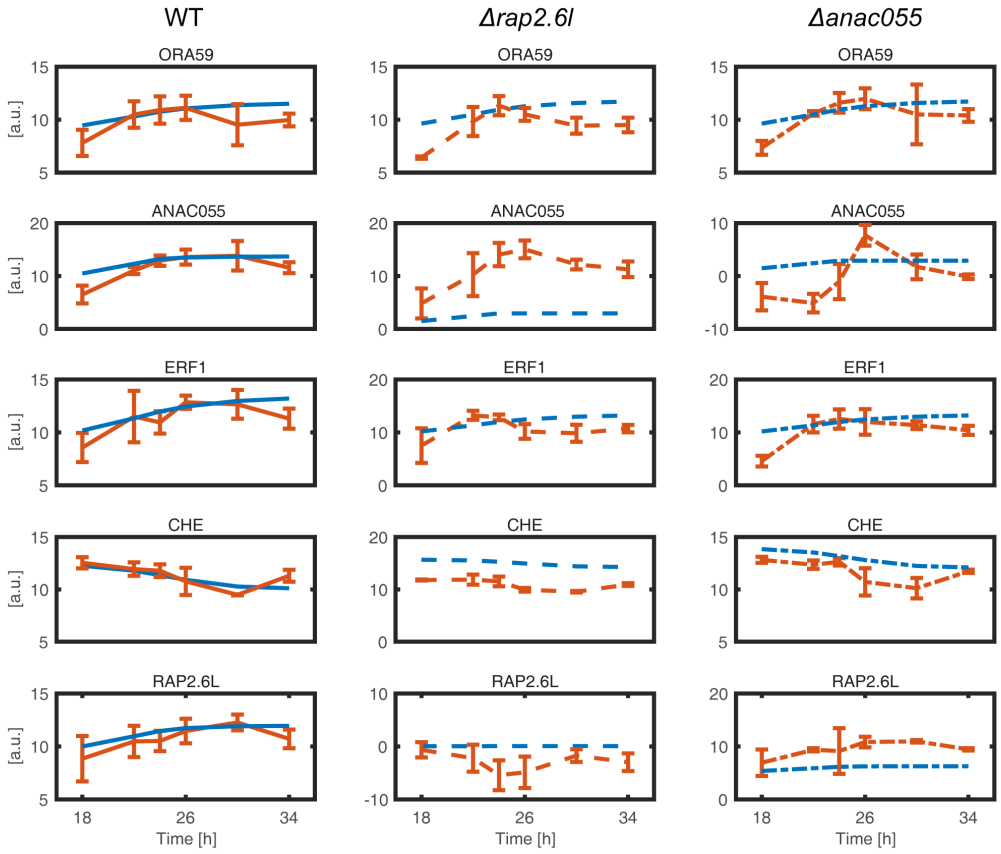


Figure 3.7: Validation of the 9GRN linear model against an independent experimental dataset generated using the Nanostring gene expression assay system (Kulkarni, 2011). Gene levels of *ORA59, ANAC055, at-ERF1, CHE* and *RAP2.6L* were measured in wild-type Arabidopsis (the first column of plots) and Arabidopsis where either *RAP2.6L* (the second column of plots) or *ANAC055* (the third column of plots) were knocked out using T-DNA insertions. The Arabidopsis plants were infected with *Botrytis* and the Nanostring measurements were taken at 18, 22, 24, 26 and 32 hours post infection (hpi). The experimental data is shown in red and the linear model simulations are shown in blue.

3.4.5 Design of a feedback controller for perturbation mitigation.

As outlined above, our control objective is to employ feedback to prevent the reduction in *CHE* levels when the plant is subjected to pathogen attack. There are several frameworks available for designing genetic controllers.(Ang *et al.* (2010), Harris *et al.* (2015), Ang and McMillen (2013)). In Ang *et al.* (2010) and Ang and McMillen (2013), the authors proposed and extended a framework for implementing an integral controller using a negative feedback of a two-promoter gene network. In Harris *et al.* (2015), the authors analyzed the dynamics of gene regulation using frequency domain tools

from control theory and proposed the implementation of a genetic phase lag controller. Here, we based our design on the framework proposed in Harris *et al.* (2015), where the proposed genetic controller is made up of a combination of genes and the regulatory relationships between them. In Harris *et al.* (2015), these gene regulations are modeled using nonlinear Michaelis-Menten type functions and these functions are then linearised such that the controller design and analysis can be done using standard frequency domain methods. In this study, since we have used a linear model to describe the 9GRN, we also model the gene regulations in the controller using linear functions.

Figure 3.8a shows the genetic circuit diagram of the proposed feedback controller. The controller architecture is modified from the framework suggested in Harris *et al.* (2015), whereby for the purposes of implementation in plants we replace the protease degradation component with a transcriptional inhibitor component. The modified circuit contains three genes and their associated proteins: genes X, Y and E giving proteins X, Y and E.

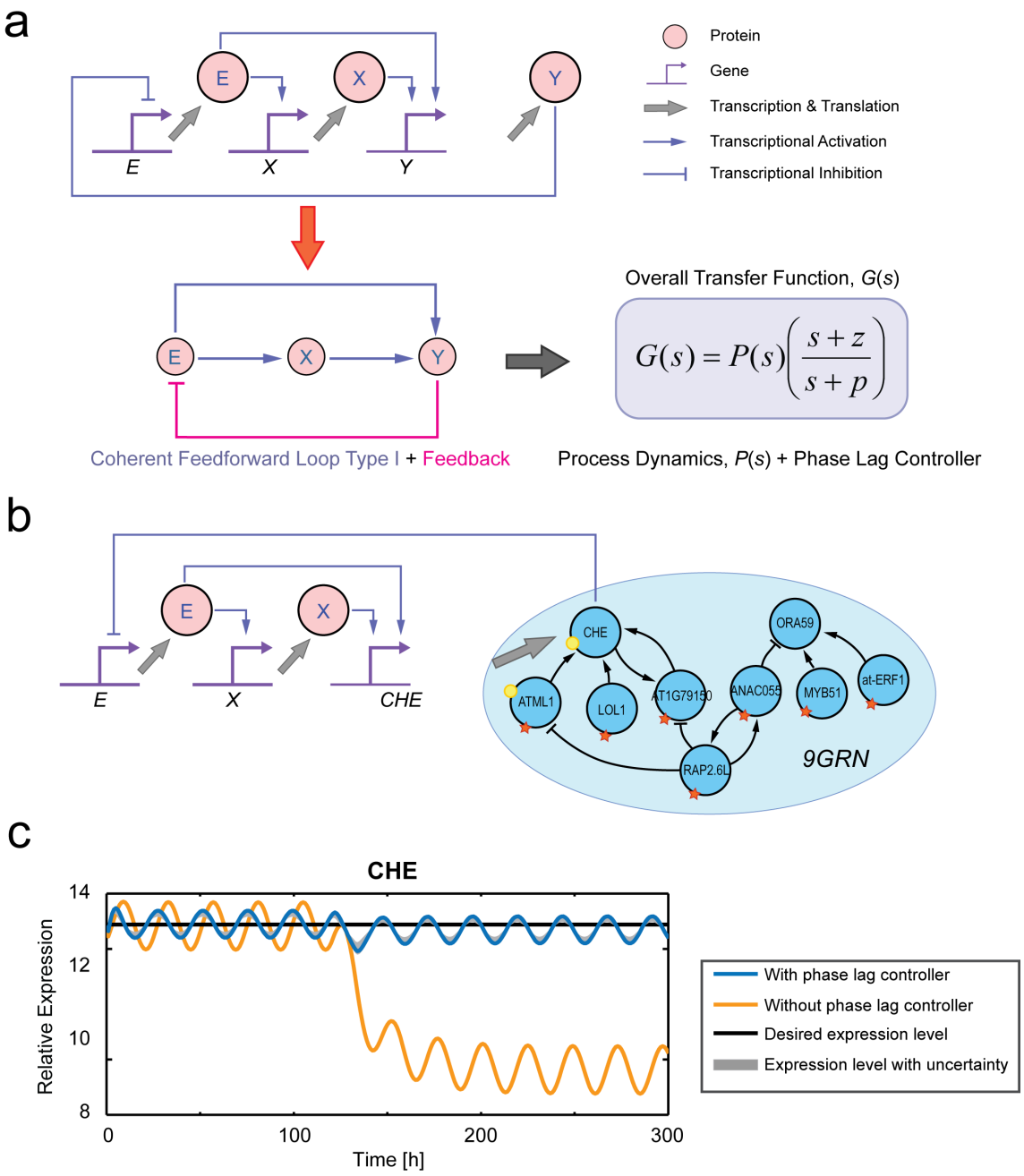


Figure 3.8: Perturbation mitigation using a genetic phase lag controller. (a) Genetic circuit of the proposed controller. X is the output of the controller, Y is the output of the process and E computes the error signal. This genetic circuit is equivalent to a coherent feedforward loop type-I with feedback network motif that yields the transfer function of a phase lag controller plus process dynamics. (b) Implementation of the phase lag controller motif for perturbation mitigation in the 9GRN. (c) Simulation of the results of phase lag controller in mitigating perturbation in the 9GRN.

Figure 3.8: *(previous page)*: The solid black line is the desired average expression of *CHE*, the solid yellow line is the expression of *CHE* during infection with *B. cinerea* without any control action, and the solid blue lines represent gene expression during infection with *B. cinerea* with control action. The gray shaded regions represent the expression level with uncertainty obtained through Monte Carlo simulation. In our simulations, the parameter values for the phase lag controller are $\alpha_{X,E} = 3.00, \alpha_{Y,X} = 5.00, \alpha_{Y,E} = 5.00, \beta_X = 0.026$, while the parameter values for the error computation are $b_{S,E} = 6.21$ and $\gamma = \beta E = 0.50$.

Let X denote an arbitrary gene that can be regulated by E, and its translated protein X denotes the TF that can regulate the output gene, Y, whose levels we ultimately want to control. E denotes the protein whose function is to regulate gene X and calculate the error signal. Here the error signal is the difference between the desired reference level and the output signal Y (see Section S1 of the Supporting Information). The ODE for the regulation of X by E is given by:

$$\frac{dX}{dt} = \alpha_{X,E}E - \beta_X X + b_{S,X} \tag{3.2}$$

Here, α, β and b_S represent production rate, degradation rate and basal expression level respectively. With Y being the output of the process that we want to control, then the ODE describing the regulation of Y by X and E can be written as:

$$\frac{dY}{dt} = \alpha_{Y,X}X + \alpha_{Y,E}E - \beta_Y Y + b_{S,Y} \tag{3.3}$$

Taking Laplace Transforms of Equations 3.2 and 3.3, and after some algebraic manipulation, we obtain the following transfer function:

$$\frac{Y(s)}{E(s)} = \left(\frac{s + \beta_X + (\alpha_{X,E}\alpha_{Y,X}/\alpha_{Y,E})}{s + \beta_X}\right)\left(\frac{\alpha_{Y,E}}{s + \beta_Y}\right) \tag{3.4}$$

Equation 3.4 is the open-loop transfer function from E to Y. In control theory, an open-loop transfer function is defined as the ratio of the output signal to the input signal in the absence of feedback and it is usually composed of the product of the transfer functions of the controller and the process. In a transfer function, the solutions making the numerator to zero are called the zeros of the system while the solutions making the denominator zero are called the poles of the system. Since Y is the output of the process, its transfer function is given by $\alpha_{Y,E}/s + \beta_Y$. Thus, the transfer function of the controller is then given by $(s + \beta_X + (\alpha_{X,E}\alpha_{Y,X}/\alpha_{Y,E}))/(s + \beta_X)$, where the zeros and poles of the controller are $-((\beta_X + \alpha_{X,E}\alpha_{Y,X})/\alpha_{Y,E})$ and $p = -\beta_X$, respectively. Since $|p| < |z|$, we obtain a phase lag controller. In control engineering, phase lag controllers are commonly used to improve disturbance rejection and reduce steady-state error (Franklin *et al.*, 2015), and thus they are well suited to our control objective of achieving perturbation mitigation. Interestingly, based on the schematic diagram of the phase lag controller as shown in Figure 3.8a, we note that this controller structure

is equivalent to a coherent feedforward loop type-I network motif (Milo *et al.* (2002), Alon (2007)), but with an added feedback loop. The role of this network motif in natural biological systems has been subjected to extensive studies and one of its key roles includes perturbation attenuation (Ma *et al.* (2009), Huang and Ren (2017)). Only certain network motifs lead to perfect adaptation.

We illustrate here in simulation the use of the genetic phase lag controller in mitigating the perturbation affecting *CHE* in the 9GRN. The configuration for perturbation mitigation using the genetic phase lag controller is shown in Figure 3.8b. In 9GRN, the output gene Y is *CHE* and the feedback is delivered by *CHE*'s transcriptional repressor activity on gene E. As with standard perturbation mitigation strategies in feedback control theory, when a perturbation causes the output level to deviate from its desired level, the controller upon detecting this deviation will react in order to restore the output to its desired level.

As *CHE* is a circadian gene, its expression level is not constant but oscillatory (*ATML1* is also light regulated). In the absence of perturbations, the *CHE* expression levels oscillate around the relative expression value of 12.44 (black line in Figure 3.8c). In our simulations, the perturbation (*B. cinerea* inoculation) is introduced at 120 hours. Upon infection by *B. cinerea*, the average expression level of *CHE* drops from 12.44 to 9.77 as indicated by the yellow solid line in Figure 3.8c. The phase lag controller upon detecting this drop in the expression level of *CHE* should exert an appropriate control action to restore the level of *CHE* to its original level. When the phase lag controller is implemented (blue solid line), the controller almost completely attenuates the effect of the perturbation with the level of *CHE* oscillating around 12.33. Moreover, this control strategy is shown to be robust against variation in model and controller parameters through a Monte Carlo simulation (see Methods section), where we randomly varied the parameters within 20% of their nominal values.

It is known from control theory that to exactly restore the output to the desired reference level after a step disturbance requires an integral-type controller (Ogata, 2009). In terms of the controller transfer function, an integral-type controller has a pole at $s = 0$. The transfer function of the phase lag controller given in Equation 3.4 has a pole at $s = -\beta X$, and therefore the slower the degradation rate for X (which corresponds to a longer mRNA half-life), the more closely the controller will implement an integral-type control action that exactly restores the output to the desired reference level after a disturbance. In Arabidopsis, the longest half-life reported for mRNAs is approximately 26 hours (Narsai *et al.*, 2007), which corresponds to a degradation rate of 0.026 /hour (calculated using the standard equation for exponential decay, $\beta = \ln(2)/T$) and therefore we have used this value in our simulations (blue solid line in Figure 3.8c). Full details of all the equations and parameter values underlying the simulations shown in Figure 3.8c can be found in Section 3.6.6.

3.4.6 Controller implementation using regulatory network rewiring.

The direct implementation of the proposed controller in Arabidopsis presents a number of challenges, largely due to the choice of TFs for E and X and associated

binding sequences. In Harris *et al.* (2015), the suggested genes for E and X are *RhaS* (E in Figure 3.8a) and *XylS* (X in Figure 3.8a). RhaS activates the production of *XylS* and *CHE* through a coherent feedforward loop, and XylS also acts as a regulator for the production of *CHE*. However, orthogonal TFs may not function in plants whilst using endogenous TFs is likely to have unintended consequences on other processes.

To get around these problems we propose an alternative approach for implementing the proposed controller, based on network rewiring. As shown in Figure 3.8a, the structure of a genetic phase lag controller is composed of a coherent feedforward loop type I motif with negative feedback. Thus, if we are able to realise this network motif through the rewiring of the 9GRN, we can obtain a genetic phase lag controller without the need to introduce new non-endogenous genetic circuitry.

For the 9GRN network shown in Figure 3.4, there are 46 potential rewiring combinations that can realise the network motif of a phase lag controller. Three of these are illustrated in Figure 3.9. However, not all genes within the 9GRN can be used in the rewiring exercise, due to functional constraints. Genes *ATML1*, *LOL1* and *AT1G79150* are not suitable for rewiring, as during *B. cinerea* infection, their expression levels decrease and to use them as part of the positive regulation of the network motif would lead to further decrease in the level of *CHE*. Another constraint is due to the gene *at-ERF1*, which is a negative regulator of plant defence (see Figure 3.3), and hence we would not wish to increase its expression further. Using *at-ERF1* as part of the positive regulation of the network motif, however, would lead to an increase in its expression. In addition, the gene *ORA59* is a positive regulator of defence, so decreasing its levels would negatively affect the defence response to *B. cinerea*. The gene *RAP2.6L* is highly responsive to stress hormones (Hickman *et al.* (2017), Krishnaswamy *et al.* (2011)) and while its involvement in infection with *B. cinerea* has not been conclusively proven, we have also chosen to discard rewiring combinations that decrease its levels. Taking these constraints into account, we are left with 11 possible rewiring combinations.

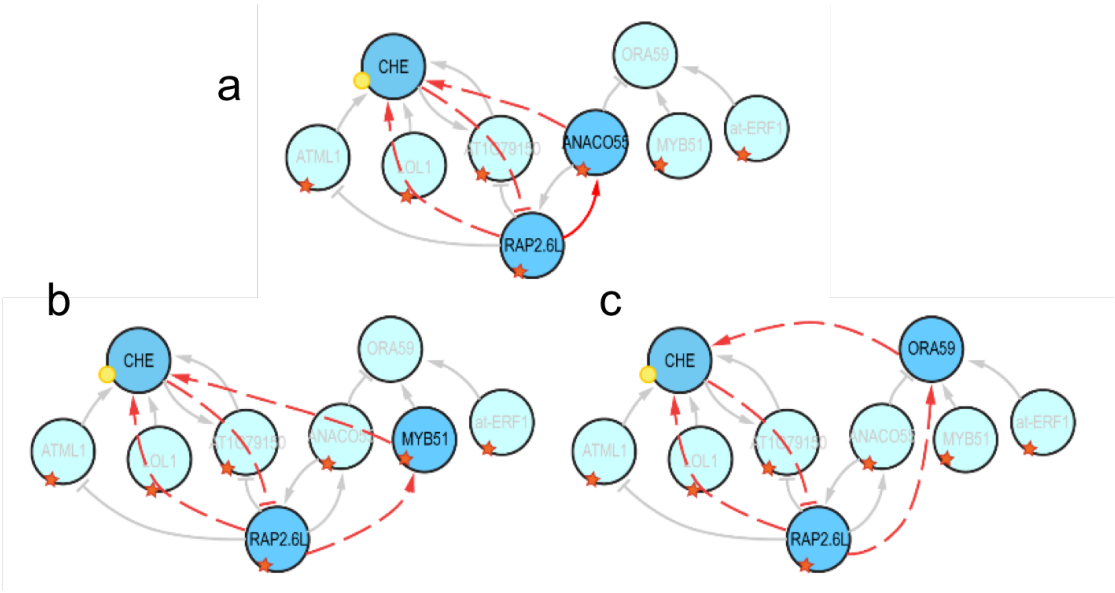


Figure 3.9: Three different possible rewiring combinations in 9GRN to realise the network motif of a phase lag controller. Red solid line: original regulation. Red dotted line: rewired regulation. (a) The equivalent Gene E is *RAP2.6L* and the equivalent Gene X is *ANAC0555*. (b) The equivalent gene E is *RAP2.6L* and the equivalent Gene X is *MYB51*. (c) The equivalent gene E is *RAP2.6L* and the equivalent Gene X is *ORA59*.

However, when *RAP2.6L* is used as the equivalent of either gene E or gene X it usually leads to a further decrease in the level of CHE. This is because RAP2.6L will increase its negative regulation of *ATML1* and *AT1G79150*, which subsequently leads to further decrease of the level of *CHE* instead of increasing it. The rewiring strategy that requires the least amount of experimental modification involves the pathway from *MYB51* (E) to *ORA59* (X) to *CHE* (Y). Note that we have included the equivalent function of the genetic phase lag controller in brackets.

Figure 3.10a shows the rewiring configuration using the pathway from *MYB51* to *ORA59* to *CHE*. To realise the required network motif, CHE must inhibit expression of *MYB51*, and MYB51 and ORA59 must activate *CHE* expression. Implementing this in simulation, with the perturbation introduced at time 120 hours, we notice only a small recovery in the expression level of *CHE* from around 9.77 to 10.31 after the perturbation (Figure 3.10a). Why is the increase in the level of *CHE* small given that we have implemented a phase lag controller through network rewiring? From Equation 3.4, we note that the pole of the phase lag controller is given by the degradation rate of X, and in this network motif, this corresponds to the degradation rate of *ORA59*. From Table 3, the value of the degradation rate of *ORA59* is 38.0062, which corresponds to placing the pole at $s = -38.0062$. From our previous discussion, it is desirable to have the pole of the controller to be as close to 0 in order for the controller to restore the output to its desired reference level. To move the pole associated with *ORA59* closer to

0, we use positive autoregulation (Ang *et al.* (2010), Harris *et al.* (2015), Drengstig *et al.* (2012)), i.e. we further rewire the network so that *ORA59* activates itself. As expected, with the addition of auto-activation of *ORA59*, we observe that the expression level of *CHE* begins to show a significant increase at around 140 hours. However, instead of returning to its original level, it increases by an extra 15% compared to its original value (Figure 3.10b). A detailed look at the plot of *MYB51* reveals that the error computed by *MYB51* is higher than expected. The reason for the incorrect error computation is that there is unmodeled regulation affecting *MYB51* (see Section 3.6.6. As a result, the controller 'sees' a larger error than actually exists, and thus exerts a higher control action to mitigate this error, resulting in the observed further increase in the expression level of *CHE*.

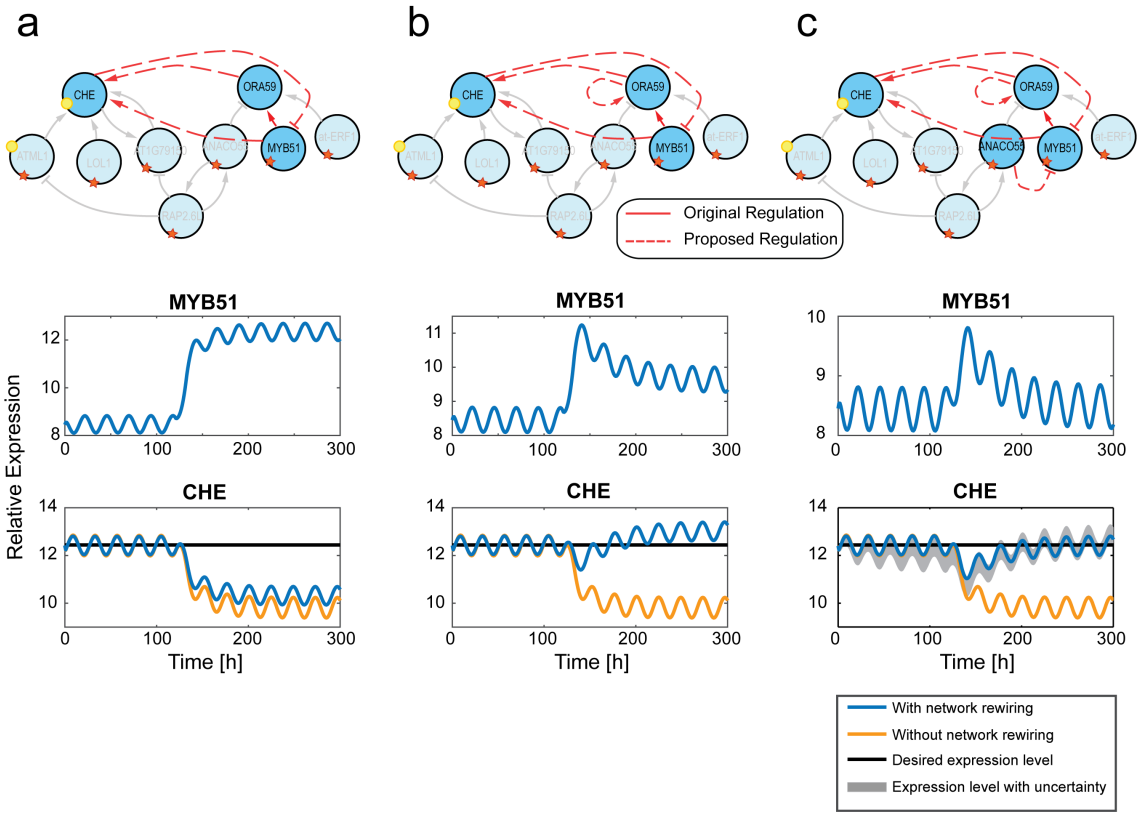


Figure 3.10: Simulation results for genes in the 9GRN with proposed network rewiring. Black line: reference value, Blue line: gene expression level in response to *B. cinerea* infection after rewiring. Yellow line: gene expression level in response to *B. cinerea* infection without network rewiring. Perturbation (inoculation) is given at time 120 hours. (a) Rewiring a controller by adding activation of *CHE* by MYB51 and ORA59 and inhibition of *MYB51* expression by CHE. (b) Addition of positive autoregulation to ORA59. (C) Addition of feedforward component; inhibition of *MYB51* by ANAC055. The gray shaded regions represent the expression level with uncertainty obtained through Monte Carlo simulation.

To address this issue, a mechanism to negate the effect of unmodeled regulation on *MYB51* is required. This can be achieved by rewiring another gene, for example *ANAC055*, to regulate *MYB51*. As the negation is independent of the process output, this is equivalent to using a feedforward controller. With the addition of autoregulation and feedforward control, the simulation results in Figure 3.10c show that the phase lag controller implemented via rewiring is now able to significantly attenuate the effect of the perturbation on *CHE* and return it to its original expression level. Additionally, the Monte Carlo simulations (see Methods section) show that the proposed strategy is robust against parameter variations. The details of the equations and parameter values underlying the simulations shown in Figure 3.10c can be found in Section 3.6.6.

3.5 Discussion

We have presented a novel strategy, based on the use of feedback control, for mitigating the effects of pathogen attack on plant gene regulatory networks, and demonstrated via simulation the ability of this approach to restore the levels of *CHE*, a key defence gene in Arabidopsis, after infection by *B. cinerea*. The use of simple rewiring such as negative autoregulation of *CHE* and direct regulation from ANAC055 was found to be insufficient for restoring the level of *CHE* (modelling results not shown), therefore we employed a coherent feedforward type I motif with negative feedback. In order to develop the strategy, we employed system identification techniques to build and validate a new dynamical model of the infected gene regulatory sub-network that accurately predicts the type of regulation between each node of the network. Then, using this model, we designed perturbation mitigation strategies using feedback control theory. In the proposed approach, we applied a combination of two positive and one negative regulatory interactions to implement genetic circuitry realising a phase lag controller. Phase lag controllers are widely used in engineering systems to reduce the effects of disturbances on system performance, and have been proposed as a useful motif for implementing synthetic biological control systems (Harris *et al.*, 2015). To date, however, practical strategies for implementing such controllers in vivo remain to be elucidated. Here, based on the observation that this control architecture resembles a coherent feedforward loop type-I with negative feedback, we propose a novel controller implementation strategy based on identifying groups of genes within the 9GRN whose regulation can be rewired to realise this network motif. Within the 9GRN, rewiring the pathway from *MYB51* to *ORA59* to *CHE* was shown to provide the most straightforward implementation of the phase lag controller. When suitably augmented with rewired autoregulation and feedforward components, this implementation of the controller was shown to deliver almost perfect perturbation mitigation without the need for any non-endogenous synthetic circuitry.

The regulatory network rewiring described above can be carried out experimentally through the insertion of constructs expressing the desired TF from the appropriate promoter region or TF binding sites combined with a minimal promoter sequence. Given that there are multiple TF binding sites in a typical promoter sequence (see e.g. Hickman *et al.* (2013)), it is preferable to use specific TF binding regions. For the rewiring we propose for the 9GRN, regulation of *MYB51* by CHE, *CHE* by ORA59, and *ORA59* regulation of its own expression, could be achieved using specific promoter regions that have been shown to confer the necessary regulation to drive expression of copies of the target TF coding sequence. CHE binds to the promoter of its target gene *CCA1* at the sequence GGTCCCAC (Pruneda-Paz *et al.*, 2009). Both the region -363 to -192 bp of the *CCA1* promoter encompassing this sequence and a trimer of the *CHE* binding sequence have been shown to be bound by CHE (Pruneda-Paz *et al.*, 2009). *ORA59* binds to two GCC boxes (GCCGCC and GCAGCCGCT) in the PDF1.2 promoter and a tetramer of one of these boxes is sufficient for *ORA59* activation of expression (Zarei *et al.*, 2011). The other regulatory edges required for rewiring (MYB51 activation of *CHE* expression and ANAC055 inhibition of *MYB51*) would currently require using the full length pro-

moter sequences and potentially fusion of transcriptional repression domains. Rewiring using full-length promoter sequences could be achieved relatively quickly (1-2 years) and methods to insert multiple gene constructs into Arabidopsis are available (for example, Golden gate cloning (Engler *et al.*, 2009)). However, the site of insertion of the necessary transgenic constructs (which is not controlled) may also influence resulting levels of expression and hence further optimization/selection of lines with appropriate levels will no doubt be necessary. Rewiring could also be used in conjunction with protein engineering to alter the binding specificity or strength of the transcription factor to cis-regulatory regions. Rhodius *et al.* (2013) generated chimeric proteins by shuffling DNA-binding domains which then bound to new targets.

The main premise of this chapter is to demonstrate the potential application of the phase lag controller motif in preventing pathogen-induced perturbations of gene expression. Our 9GRN model is used to demonstrate how a phase lag controller could function. External feedback controllers have successfully been introduced to control levels of specific genes. Milias-Argeitis *et al.* (2011) have used real-time Fluorescence Activated Cell Sorter measurements of a fluorescent protein, YFP, in yeast as input to a computer that calculates its concentration. The light-responsive proteins PhyB - fused to the GAL4 DNA-binding domain - and PIF3 - fused to the GAL4 activation domain - were controlled by red and far-red pulses of light to switch on and off YFP production. The promoter of YFP contained GAL4 binding sites. The computer was then used to achieve the desired fluorescence set-point by delivering pulses of light to the yeast. A similar system based on Phy and PIF proteins was used by Toettcher *et al.* (2011) to control the membrane concentration of PIP_3 levels by spiking in serums that lead to its production or degradation. While the actuator is controlled by a computer rather than the cell, an antithetic integral feedback motif has recently been introduced which contains all aspects of the controller within two plasmids inside an *E. coli* cell and constructed *in vivo* (Briat *et al.*, 2016a; Lillacci *et al.*, 2017). The design consists of 4 species: the population of interest, X_l, the actuated species X_1, the control input species Z_1 and the sensing species Z_2. Z_1 influences the rate of production of X_1, which influences the rate of production of the regulated species X_l. X_l, in turn, influences the levels of Z_2 which combines with Z_1 in an annihilation reaction which eliminates both species (Briat *et al.*, 2016a). This negative feedback loop was implemented in *E. coli* using GFP/AraC as the regulated species X_l. Arabinose and HSL are used to move the set-point (Lillacci *et al.*, 2017). *Bacillus subtilis* σ and anti-σ factors are the species that annihilate and the system also requires a a protease, mflon, to degrade the GFP and AraC molecules. As our aim was to model and regulate everything at the gene level, we did not include the use of protease as was done in (Lillacci *et al.*, 2017) and (Harris *et al.*, 2015).

We have provided some evidence for edges in our network, but the presence of additional edges we have not modelled or false positive edges could have a significant impact on the performance of this controller in vivo. Strategies such as DAP-seq (O'Malley *et al.*, 2016) are making significant improvements in our knowledge of plant TF-promoter interactions but, particularly given the expansion of TF families in plants (Shiu *et al.*,

2005), greater mapping of plant gene regulatory networks under multiple environmental and developmental conditions will be necessary to drive successful plant synthetic biology strategies. Clearly, in all non-orthogonal rewiring strategies the new edges may have unintended consequences on plant physiology through changing regulatory interactions. In our *in silico* implementation, the controller is only triggered by a significant reduction in *CHE* levels (such as that driven by pathogen infection, not the daily circadian oscillations), and levels of *CHE* and *MYB51* quickly return to normal. The intention of the controller is to maintain oscillating expression levels of *CHE* given its key role in the circadian clock (regulating CCA1, a core transcriptional regulator (Pruneda-Paz *et al.*, 2009)). We also ensured that expression of positive regulators of defence was not compromised. However, the genes with new rewired links (*MYB51*, *ORA59* and *ANAC055*) are involved in response to other environmental conditions. For example, *MYB51* promotes expression of indolic glucosinolate biosynthetic genes (Gigolashvili *et al.*, 2007) in response to mechanical stimuli and *ANAC055* is induced by drought, salt and abscisic acid stress (Tran *et al.*, 2004). Changes in the expression of these genes due to other stimuli could prevent the controller from operating during *B. cinerea* infection (for example, induction of *ANAC055* would lower levels of *MYB51* and indicate a lower level of error to the controller, leading to the controller not reacting properly). Our controller is not designed to handle more than one perturbation (environmentally induced shift in gene expression) and this limitation raises another key challenge in plant systems biology. The development of novel approaches to model and simulate dynamic networks of sufficient size to capture environmental stress cross-talk will significantly improve our ability to rationally engineer stress resilient plants.

3.6 Methods

3.6.1 Transgenic Arabidopsis line.

The *CHE* T-DNA insertion line, SALK_143403c, was obtained from the SALK collection (Alonso *et al.*, 2003) and confirmed to be homozygous. Expression of *CHE* in this line is significantly reduced compared to wildtype (Col-0) (Pruneda-Paz *et al.*, 2009). The coding region of *at-ERF1* was cloned into the pB7WG2 vector (Karimi *et al.*, 2002) and stable transgenic Arabidopsis lines generated in a Col-4 background. The coding region of *at-ERF1* was cloned into the pB7WG2 vector (Karimi *et al.*, 2002) and stable transgenic Arabidopsis lines generated in a Col-4 background.

3.6.2 Infection assay

B. cinerea infection assay was carried out as described in Section 5.6.13.

3.6.3 Yeast-1-Hybrid assay

Yeast-1-Hybrid assays were performed as previously described (Hickman *et al.*, 2013). Three overlapping promoter regions (of approximately 400 bp) spanning 800

Gene	Forward Primer	Reverse Primer
AT1G06160	AAAAAAGCAGGCTTCGTGCAATTGATCAC TATATTAGTTGAACTG	CAAGAAAGCTGGGTCGTGTCTAAGTGGCACT AAGTTTGGG
AT1G06160	AAAAAAGCAGGCTTCCCGCCTTAGTTTCTG ACAGAGTTTCGACTC	CAAGAAAGCTGGGTCGAGTGTATGACGTAC GGCGGCGTATTCCCG
AT1G06160	AAAAAAGCAGGCTTCCTGTTCTGTCGAGTT GTTGCTTGTTGAGCC	CAAGAAAGCTGGGTCTGTGGGCAAAATAGG TCAAACATGCGGC
AT3G15500	GGGGGAATTCATAAGAGGAGGTACAGTC ACACA	GGGGAAGCTTACGCGTCGAAGCTCTGCTACT CGTGTATGTAT
AT3G15500	GGCCGAATTCATCCCATCATTCACTTACAC	GGGGAAGCTTACGCGTGATCAATTAGAGCG TCGTGATTTATGC
AT3G15500	GGGGGAATTCGTTTGTTGTTTGTCCCTCTC TCTGA	GGGGAAGCTTACGCGTTGAGTTACATAACAG TGACAATCTACGA
AT3G15500	GGGGGAATTCGAGAAGCGTGTTTGTGTTA TACGGACTTA	GGGGAAGCTTACGCGTTGTGTCTATTGGTTG AGTTAGGC

Table 3.1: Primer sequences for cloning promoter fragments for Y1H.

to 1200 bp upstream of the transcription start site were used as bait for transcription factors fused to a GAL4 activation domain in pDEST22 (Invitrogen). Yeast strain AH109 (Clontech) was transformed with these individual TF clones. The promoter fragments were amplified using two-step PCR and cloned into a pDonrZeo vector (Invitrogen) using Gateway cloning. Yeast strain Y187 (Clontech) was transformed with the individual vectors to create the bait strain. The promoter strain was spotted onto YPDA (yeast, peptone, dextrose, adenine) plates, overlaid with the TF strain, and incubated for 24 hours at 30°C. The diploid cells were replica plated onto selective plates and incubated overnight. This was followed by replica-cleaning and incubation for 4 days, after which growth was scored. Each interaction was tested twice. Primer sequences for the promoter fragments are given in Table 3.1.

3.6.4 Accession numbers

Arabidopsis gene names and AGI locus codes referred to in this article are shown in Table 3.2.

Table 3.2: Associated AGI to Arabidopsis gene names.

ORA59	AT1G06160
MYB51	AT1G18570
LOL1	AT1G32540
AT1G79150	AT1G79150
ANAC055	AT3G15500
at-ERF1	AT4G17500
ATML1	AT4G21750
CHE	AT5G08330
RAP2.6L	AT5G13330

3.6.5 Generating the TF network.

An Arabidopsis TF list was generated by combining lists from ThaleMine (Krishnakumar *et al.*, 2015), DATF (Guo *et al.*, 2005), Pruneda-Paz *et al.* (2014), and homology searches using DNA binding domains, followed by manual curation of genes only identified in one list. The final list of 2,534 genes is included in Appendix B. The list of Arabidopsis genes differentially expressed during *B. cinerea* infection was obtained from Windram *et al.* (2012), which included 883 differentially expressed TFs.

3.6.6 Generating a dynamic model of the *CHE* regulatory sub-network.

Traditionally, a model of a gene regulatory network comprises both transcription and translation mechanisms. However, in our case, given that only mRNA accumulation time-series data are available (Windram *et al.*, 2012), the following two assumptions are made in building the 9GRN model. Firstly, the translation of the protein from mRNA follows a linear relationship (see section 1.6.1.2) and secondly, the behavior of the translated protein follows its mRNA closely. With these two assumptions, we can group together the protein translation rate with the mRNA transcription rate resulting in the entire 9GRN being modeled using only mRNA data. Based on the above assumptions, the model of the 9GRN shown in Figure 3.4 can be described by the ODEs in Equation 3.5.

$$
\begin{aligned}
\frac{dN_{ORA}}{dt} &= \alpha_{ORA,1}N_{MYB} + \alpha_{ORA,2}N_{ANA} + \alpha_{ORA,3}N_{ERF} + \beta_{ORA}N_{ORA} + b_{S,ORA},\\
\frac{dN_{MYB}}{dt} &= \beta_{MYB}N_{MYB} + b_{S,MYB} + c_{MYB}W,\\
\frac{dN_{LOL}}{dt} &= \beta_{LOL}N_{LOL} + b_{S,LOL} + c_{LOL}W,\\
\frac{dN_{AT1}}{dt} &= \alpha_{AT1,1}N_{CHE} + \alpha_{AT1,2}N_{RAP} + \beta_{AT1}N_{AT1} + b_{S,AT1} + c_{AT1}W,\\
\frac{dN_{ANA}}{dt} &= \alpha_{ANA,1}N_{RAP} + \beta_{ANA}N_{ANA} + b_{S,ANA} + c_{ANA}W,\\
\frac{dN_{ERF}}{dt} &= \beta_{ERF}N_{ERF} + b_{S,ERF} + c_{ERF}W,\\
\frac{dn_{ATM}}{dt} &= \alpha_{ATM,1}N_{RAP} + \beta_{ATM}N_{ATM} + b_{S,ATM} + c_{ATM}W + \gamma_{ATM}L,\\
\frac{dN_{CHE}}{dt} &= \alpha_{CHE,1}N_{LOL} + \alpha_{CHE,2}N_{AT1} + \alpha_{CHE,3}N_{ATM} + \beta_{CHE}N_{CHE} + b_{S,CHE} + \gamma_{CHE}L,\\
\frac{dN_{RAP}}{dt} &= \alpha_{RAP,1}N_{ANA} + \beta_{RAP}N_{RAP} + b_{S,RAP} + c_{RAP}W.
\end{aligned}
\tag{3.5}
$$

The ODE model of the original GRN9, where $\alpha_{i,j} \in (-\infty, +\infty)$, $\beta_i < 0$, $\gamma_{CHE} > 0$, $b_{S,i} \in (-\infty, +\infty)$, and $c_i \in (-\infty, +\infty)$ are the unknown parameters that represent the production rate, degradation rate, scaled light effect, basal level and effect of the unmodeled regulation, respectively, with i and j denoting the appropriate indices describing the parameters given in Table 3.3. N_i represents the gene. W represents the effect of the unmodeled regulation (e.g. direct regulation as a result of *B. cinerea* infection, noise and other regulations not identified by the network inference algorithms), where $W = 0$ (resp. $W = 1$) is used when the effect is absent (respectively present).

In the experiments from which our data were generated (Windram *et al.*, 2012), the time-series data from the control and infected experiments are treated as a continuous dataset where the infection starts at the halfway point, i.e., time 48 hours. Thus, the transition of W from 0 to 1 is not modeled as an instantaneous change but as a gradual increase. L represents the effect of light and *CHE* follows a sinusoidal rhythm as previously described (Pruneda-Paz *et al.*, 2009). The values of the model parameters were estimated from the available mRNA time-series data using a nonlinear least squares algorithm and the estimated parameters are given in Table 3.3.

All the simulations of the ODE models, phase genetic controller and network rewiring are done using MATLAB built-in solver ode45, and the initial condition for each gene to solve the ODE is the first data point of the mRNA time-series for each respective gene. For the simulation using the genetic phase lag controller, the initial conditions for solving the ODEs for X and E are set to 0.

Gene Name	Values
ORA59	$\alpha_{ORA,1} = 14.3800$, $\alpha_{ORA,2} = -0.7359$, $\alpha_{ORA,3} = 21.5714$, $\beta_{ORA} = -38.0062$, $b_{S,ORA} = 15.2355$
MYB51	$\beta_{MYB} = -0.6658$, $b_{MYB} = 5.6277$, $c_{MYB} = 1.1890$
LOL1	$\beta_{LOL} = -0.0485$, $b_{S,LOL} = 0.4874$, $c_{LOL} = -0.1241$
AT1G79150	$\alpha_{AT1,1} = 0.7577$, $\alpha_{AT1,2} = -0.7408$, $\beta_{AT1} = -2.4088$, $b_{S,AT1} = 23.863$, $c_{AT1} = 0.91809$
ANAC055	$\alpha_{ANA,1} = 25.6935$, $\beta_{ANA} = -28.4685$, $b_{S,ANA} = 0.0517$, $c_{ANA} = 82.5415$
at-ERF1	$\beta_{ERF} = -0.2051$, $b_{S,ERF} = 1.8699$, $c_{ERF} = 0.8735$
ATML1	$\alpha_{ATM,1} = -0.7945$, $\beta_{ATM} = -1.1142$, $b_{S,ATM} = 19.3684$, $c_{ATM} = 0.0040$, $\gamma_{ATM} = 0.5000$
CHE	$\alpha_{CHE,1} = 24.5024$, $\alpha_{CHE,2} = 3.3801$, $\alpha_{CHE,3} = 17.6771$, $\beta_{CHE} = -40.1258$, $b_{S,CHE} = 3.7167$, $\gamma_{CHE} = 16.8001$
RAP2.6L	$\alpha_{RAP,1} = 0.4186$, $\beta_{RAP} = -0.7933$, $b_{S,RAP} = 3.7046$, $c_{RAP} = 0.0045$

Table 3.3: Estimated parameters of the linear model

The ODE representation of perturbation mitigation using genetic phase lag controller to obtain the simulation results shown in Figure 3.8c is given in Table 3.4.

The ODE representation of perturbation mitigation using network rewiring to obtain the simulation results shown in Figure 3.10 is given in Table 3.5.

The parameters of the rewired network used to obtain simulation shown in Figure 3.10 are: $\alpha_{ORA,3} = 37.890$, $b_{S,ORA,RW1} = -327.388$, $\alpha_{MYB,1} = -0.211$, $b_{S,ERF,RW} = 1.6995$, $b_{S,CHE,RW} = 8.267$, $\alpha_{CHE,4} = 0.030$, $\alpha CHE, 5 = 5.000$, $b_{S,ORA,RW2} = -0.259$, $b_{S,MYB,RW} = -42.261$. The subscript ${}_RW$ refers to the parameters of the rewired model.

3.6.7 Mathematical Representation of Unmodeled Regulations and Light Effect on CHE.

Some regulations (e.g. direct regulation as a result of *B. cinerea* infection, noise and other regulations not identified by the network inference algorithms) are not considered directly in the ODE model and are treated as unmodeled regulation. We denote them by W where $W = 1$ is used when the effect of the unmodeled regulation is present and $W = 0$ when the effect is absent. As the effect of *Botrytis* and unmodeled regulation on the 9GRN is not instantaneous but a gradual one, we model W with a gradual increase from $t = 48$ to 72 hours prior reaching maximum at 1:

$$W(t) = \begin{cases} 0 & 0 \leq t < 48 \\ (1/24)t - 2 & 48 \leq t \leq 72 \\ 1 & t > 72 \end{cases} \tag{3.6}$$

Since *CHE* and *ATML1* are circadian genes, these genes are affected by the light condition. In the original experiment, the light condition used was 16 hours of light and 8 hours of dark. The resulting *CHE* expression behaves in a sinusoidal manner with clear peak at between 8-10 hours after dawn. The term L denotes the effect of light is modeled as a sinusoidal signal with period of 24 hours that has its peak around 9 hours after dawn:

$$L(t) = A \sin\left(\frac{2\pi t}{T} + \phi\right) + B \tag{3.7}$$

where A is the amplitude, $T = 24$ hours, ϕ is the phase shift and B is the basal level. To generate the desired light effect for *CHE*, $A = 16.8$ while for *ATML1*, $A = 0.5$ (both obtained from parameter estimation) $B = 1$ and $\phi = \pi/6$.

3.6.8 Parameter estimation.

For the 9GRN linear model, the values of the unknown parameters are estimated from the available mRNA time-series using nonlinear least square, given by

$$\hat{\theta} = \arg\min_\theta \frac{1}{N_L} \sum_{i \in \psi} \sum_{t=1}^{N_L} \left[N_i(t) - \hat{N}_i(t, \theta) \right]^2 \tag{3.8}$$

$$\frac{dN_{ORA}}{dt} = \alpha_{ORA,1}N_{MYB} + \alpha_{ORA,2}N_{ANA} + \alpha_{ORA,3}N_{ERF} + \beta_{ORA}N_{ORA} + b_{S,ORA}$$

$$\frac{dN_{MYB}}{dt} = \beta_{MYB}N_{MYB} + b_{S,MYB} + c_{MYB}W$$

$$\frac{dN_{LOL}}{dt} = \beta_{LOL}N_{LOL} + b_{S,LOL} + c_{LOL}W$$

$$\frac{dN_{AT1}}{dt} = \alpha_{AT1,1}N_{CHE} + \alpha_{AT1,2}N_{RAP} + \beta_{AT1}N_4 + b_{S,AT1} + c_{AT1}W$$

$$\frac{dN_{ANA}}{dt} = \alpha_{NAC,1}N_{RAP} + \beta_{ANA}N_{ANA} + b_{S,NAC} + c_{NAC}W$$

$$\frac{dN_{ERF}}{dt} = \beta_{ERF}N_{ERF} + b_{S,ERF} + c_{ERF}W$$

$$\frac{dN_{ATM}}{dt} = \alpha_{ATM,1}N_{RAP} + \beta_{ATM}N_{ATM} + b_{S,ATM} + c_{ATM}W + \gamma_{ATM}L$$

$$\frac{dN_{CHE}}{dt} = \alpha_{CHE,1}N_{LOL} + \alpha_{CHE,2}N_{AT1} + a_{CHE,3}N_{ATM} + \beta_{CHE}N_{CHE} + b_{S,CHE} + \gamma_{CHE}L + U$$

$$\frac{dN_{RAP}}{dt} = \alpha_{RAP,1}N_{ANA} + \beta_{RAP}N_{RAP} + b_{S,RAP} + c_{RAP}W$$

$$\frac{dX}{dt} = \alpha_{X,1}E + \beta_X X + b_{S,X}$$

$$\frac{dE}{dt} = \beta_E E + b_{S,E} + \beta_E N_{CHE}$$

Table 3.4: The ODE model for general perturbation mitigation using a genetic phase lag controller in GRN9, where $U = \alpha_{Y,X}X + \alpha_{Y,E}E$ is the control action provided by the phase lag controller. The parameters of the phase lag controller used to obtain simulation shown in Figure 3.8c are: $\alpha_{X,1} = 3.000, \beta_X = -0.026, b_{S,X} = 0, \beta_E = -0.500, b_{S,E} = 6.210$.

$$\frac{dN_{ORA}}{dt} = \alpha_{ORA,1}N_{MYB} + \alpha_{ORA,2}N_{ANA} + \alpha_{ORA,3}N_{ERF} + \beta_{ORA}N_{ORA} + b_{S,ORA} + \left(\alpha_{ORA,3}N_{ORA} + b_{S,ORA,RW1}\right)$$

$$\frac{dN_{MYB}}{dt} = \beta_{MYB}N_{MYB} + b_{S,MYB} + c_{MYB}W + \left(\beta_{MYB}N_{CHE} + b_{S,CHE,RW}\right) + \left(\alpha_{MYB,1}N_{ANA} + b_{S,ANA,RW}\right)$$

$$\frac{dN_{LOL}}{dt} = \beta_{LOL}N_{LOL} + b_{S,LOL} + c_{LOL}W$$

$$\frac{dN_{AT1}}{dt} = \alpha_{AT1,1}N_{CHE} + \alpha_{AT1,2}N_{RAP} + \beta_{AT1}N_4 + b_{S,AT1} + c_{AT1}W$$

$$\frac{dN_{ANA}}{dt} = \alpha_{ANA,1}N_{RAP} + \beta_{ANA}N_{ANA} + b_{S,ANA} + c_{ANA}W$$

$$\frac{dN_{ERF}}{dt} = \beta_{ERF}N_{ERF} + b_{S,ERF,RW} + c_{ERF}W$$

$$\frac{dN_{ATM}}{dt} = \alpha_{ATM,1}N_{RAP} + \beta_{ATM}N_{ATM} + b_{S,ATM} + c_{ATM}W + \gamma_{ATM}L$$

$$\frac{dN_{CHE}}{dt} = \alpha_{CHE,1}N_{LOL} + \alpha_{CHE,2}N_{AT1} + a_{CHE,3}N_{ATM} + \beta_{CHE}N_{CHE} + b_{S,CHE} + \gamma_{CHE}L + \left(\alpha_{CHE,4}N_{ORA} + b_{S,ORA,RW2}\right) + \left(\alpha_{CHE,5}N_{MYB} + b_{S,MYB,RW}\right)$$

$$\frac{dN_{RAP}}{dt} = \alpha_{RAP,1}N_{ANA} + \beta_{RAP}N_{RAP} + b_{S,RAP} + c_{RAP}W$$

Table 3.5: The ODE model of the rewired GRN9 forming a phase lag controller. Equations for rewiring are shown in brackets.

Gene Name	MSE (Training)	MSE (Validation)
ORA59	0.773	1.9828
MYB51	0.391	0.6949
LOL1	0.3703	0.6582
AT1G79150	0.1829	0.3587
ANAC055	0.9889	2.3849
at-ERF1	0.4394	1.0583
ATML1	0.3746	0.6682
CHE	0.8759	1.0819
RAP2.6L	0.3452	0.6366
MSE_T	4.741	9.5245

Table 3.6: MSE values for both the training and validation data sets of the 9GRN using the NanoString data.

where $\theta = [\alpha_i, \beta_i, b_{S,i}, c_i]$ with $i \in \psi$ =[*ORA, MYB, LOL, AT1, ANA, ERF, ATM, CHE, RAP*], N_L is the length of the time-series data, $\hat{N}$ is the simulated data from the ODE equations and N is the experimental data, which are the mRNA time-series taken from Windram *et al.* (2012). There are four sets of mRNA time-series and we use the average mRNA expression from the first three sets for parameter estimation and use the fourth data set as an independent data set for validating the ODE model. Equation 3.8 is solved using MATLAB function fminsearch which uses the Nelder-Mead simplex algorithm.

As a quantitative measure of the model performance, we compute the Mean Square Error (MSE) for each gene between the experimental data and the simulated data. The MSE for each gene is computed as follows:

$$\text{MSE} = \frac{1}{N_L} \sum_{t=1}^{N_L} \left[N(t) - \hat{N}(t, \theta) \right]^2 \tag{3.9}$$

The total MSE, MSE_T is computed by summing the MSE for all nine genes in the 9GRN. Table 3.6 shows the MSE values for both the training and validation data sets.

3.6.9 Performance and robustness analysis.

To analyze the performance and robustness of the proposed strategies, we perform a Monte Carlo simulation where we randomly draw all the parameters from a uniform distribution. Then, we vary the parameters within ranges of 20%, around their nominal

values. Mathematically, we have $p(1 + \Delta P(x))$, where p denotes the model and the controller parameters, $P(x)$ is the probability distribution and $\Delta = 0.2$. Using the Chernoff bound and associated guidelines for Monte Carlo simulation, a total number of 1060 simulations is required to achieve an accuracy level of 0.05 with a confidence level of 99% (Vidyasagar (1998), Menon *et al.* (2009)).

Chapter 4

Gaussian process dynamical systems for optimisation of *in silico* and Arabidopsis gene regulatory networks

4.1 Aims

Having used ODEs for modelling small subnetworks in the previous chapter, it would be advantageous to be able to simulate rewiring scenarios for large networks of ∼100 genes, to more accurately predict the situation *in planta*. Therefore Gaussian processes (GPs) are introduced in this chapter as a method for modelling gene expression data over time and making predictions for rewirings of networks with many nodes and edges. The work in this chapter has the following aims:

- Learn a small *in silico* network (DREAM10.1) using time series data and simulate a validation dataset for this network using Gaussian process dynamical systems (GPDS) models.
- Learn a larger *in silico* network (DREAM100.1) using time series data and simulate a validation dataset for this network using GPDS models.
- Use GPDS models with a number of covariance functions and determine whether a certain covariance function is more suited to modelling gene expression data.
- Use GPDS models to simulate gene expression of DREAM10 and DREAM100 networks upon rewiring of genes - that is, replacing the regulators of a gene with the regulators of another. Compare the simulations to the 'true' rewiring expression values obtained from the GeneNetWeaver programme.
- Identify a subnetwork of the *At-Botrytis* network to model using GPDS models.

- Learn the *At-Botrytis* subnetwork using a GP prior as before and test the model on a validation dataset.
- Use the *At-Botrytis* subnetwork GP model to predict the gene expression upon rewiring, and identify the rewiring(s) that optimise the levels of genes in the network that are positive (or negative) regulators of defence.

4.2 Acknowledgements

Section 4.3.3 is based on 'Inferring gene regulatory networks from multiple datasets', a chapter in Gene Regulatory Networks: Methods and Protocols, in the series Methods in Molecular Biology published by Springer, co-authored by Christopher A. Penfold, Anastasiya Sybirna, David L. Wild and I. The rest of the chapter is written entirely by myself.

4.3 Introduction

A major challenge in systems and synthetic biology is the ability to model complex regulatory interactions, providing a compact and accurate representation of a biological system that is capable of predicting behaviour under novel conditions. Numerous modelling frameworks exist that can simulate gene regulatory network (GRN) data. Ordinary differential equations (linear or non-linear such as Michaelis-Menten-based) and stochastic differential equations are bottom-up approaches with a detailed mathematical description of the biophysical processes involved in gene regulation. In the previous chapter, linear ODEs were used to model the interactions between genes in a small Arabidopsis subnetwork and apply a feedback mechanism to control the levels of a gene that is important in the *Botrytis cinerea* disease process (see Table 3.5). ODEs are very well suited to describing small systems of genes, such as the aforementioned application, and the Arabidopsis circadian clock (Pokhilko *et al.*, 2010). However, they generally require in-depth knowledge of kinetic parameters such as the leaky transcription rate, mRNA and protein degradation rate, and mRNA and protein production rate. Some of these quantities, such as mRNA decay rates, can be measured in a high throughput manner (Narsai *et al.*, 2007) whereas others are harder to determine, especially for large numbers of mRNA/proteins. An alternative is to infer these parameters computationally, which requires large amounts of data. For instance, a rule of thumb is that 20 observations are necessary for inferring one parameter. Further complications can arise due to the noisiness of the data and the nonlinear dynamics of gene regulation (Steinacher *et al.*, 2016). Parametric models also make certain assumptions about the functions describing the data, such as linearity, so simply choosing one such model over another introduces certain inflexible rules; overfitting a model can also be an issue.

While parametric models require the parameters to capture all aspects of the data and predict future values, non-parametric models are represented in terms of other quantities derived from the data. They have infinitely many parameters (which can

be described in terms of a function instead), do not make assumptions about the relationships between variables and are therefore more flexible in their approach. One such nonparametric approach is the Gaussian process. Gaussian processes (GPs) can be viewed as a Bayesian method of specifying distributions over an unknown function, such as the function capturing the relationship between transcription factor (TF) regulators and the gene they are regulating. Bayesian methods generally provide a powerful framework for modelling, as they can represent uncertainty, and combine prior knowledge with data. Rather than ascribing a particular function to the observed measurements, nonparametric models let the data 'speak' for themselves. A Gaussian process is formally defined as a collection of random variables, any finite number of which have a joint multivariate Gaussian distributions (Rasmussen, 2004).

Gaussian processes are particularly suitable for Bayesian learning in problems involving classification, such as pitch accent prediction, named entity recognition (Altun *et al.*, 2004), patient classification from resting state fMRI (Challis *et al.*, 2015), facial expression recognition (Cheng *et al.*, 2010), and regression problems such as robot gait optimisation (Lizotte *et al.*, 2007), robot inverse dynamics with unknown nonlinearities (Nguyen-Tuong *et al.*, 2009), calibration for spectroscopic quantitative analysis (Chen *et al.*, 2007) and super-resolution imaging (He and Siu, 2011). Gaussian process regression (GPR), like any regression, infers the relationship between explanatory and dependant variables.

Gaussian process dynamical systems (GPDS) represent a class of probabilistic models able to capture a range of dynamical systems such as the 9GRN, in a scalable, nonparametric way. These approaches were initially developed to capture the dynamics seen in relatively small systems encountered in the physical sciences (Murray-Smith *et al.*, 1999; Sbarbaro and Murray-Smith, 2005). Developments by Klemm *et al.* (2008) applied GPDS to inferring small GRNs, and their utility was subsequently demonstrated in much larger systems containing hundreds of genes (Penfold and Wild, 2011).

4.3.1 Gaussian process dynamical systems

In Gaussian process dynamical systems, the output y of a function f is written as

$$y = f(\mathbf{x}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_n^2), \tag{4.1}$$

with ϵ representing noise in the observations and y representing the protein or gene levels of a particular transcription factor of interest. $\mathbf{x}$ is a vector of training inputs - such as the expression levels of regulator genes.

The function $f(\mathbf{x})$ is distributed as a Gaussian process. The description of a Gaussian process lies entirely in its covariance $k(\mathbf{x}, \mathbf{x}')$ and mean $m(\mathbf{x})$ functions.

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{4.2}$$

where $\mathbf{x}$ and $\mathbf{x}'$ represent inputs (vectors or matrices). The mean function reflects the average of all functions from the distribution evaluated at $\mathbf{x}$. For simplicity, the mean function can be mean-centred so that $m(\mathbf{x}) = 0$. Within this work, this is achieved by

calculating the Z-score for the expression levels of each gene so that the mean is 0. The covariance function calculates the dependence between $f(\mathbf{x})$ at all pairs of the input data $\mathbf{x}$ and $\mathbf{x}'$.

$$k(\mathbf{x}, \mathbf{x}')) = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))], \tag{4.3}$$

The most frequently used covariance function, or kernel, is the squared exponential covariance function defined below.

$$k(\mathbf{x}, \mathbf{x}')) = \sigma_f^2 exp\left[\frac{-(\mathbf{x} - \mathbf{x}')^2}{2l^2}\right], \tag{4.4}$$

where l, the characteristic length scale and σ_f, the standard deviation that describe the covariance function, are known as hyperparameters.

The length scale describes the overall trend of the data: shorter length scales imply that the function changes rapidly and that observations y_1 and y_2 of two distant inputs $\mathbf{x}_1$ and $\mathbf{x}_2$ do not have much effect on each other; whilst fitting a function with too short a length scale can given an exact fit of the data, it is not favoured by the marginal likelihood and corresponds to overfitting (Figure 4.1). The signal variance σ_f^2 determines how far values in the function can stray from the mean. As can be seen, the maximum allowable covariance is given by σ_f^2. The noise term is added to the equation to represent experimental error:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 exp\left[\frac{-(\mathbf{x} - \mathbf{x}')^2}{2l^2}\right] + \sigma_n^2 \delta(\mathbf{x}, \mathbf{x}'), \tag{4.5}$$

where $\sigma_n^2 \delta(\mathbf{x}, \mathbf{x}')$ is Gaussian noise with variance σ_n^2 when $\mathbf{x} = \mathbf{x}'$ and zero otherwise.

Given training data $[(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)]$, which has inputs $\mathbf{x}_1$ to $\mathbf{x}_n$ and outputs y_1 to y_n, the aim of regression is to learn the function that describes the dataset and predicts the value that y^* will take at $\mathbf{x}^*$. In this work, $\mathbf{x}_1$ to $\mathbf{x}_n$ are the expression values of the regulators of the gene of interest, and y_1 to y_n are the values of the gene of interest. These are used to predict the expression value of the gene of interest in a different condition or a future time point, y^*, given the expression values of its regulators, $\mathbf{x}^*$.

To do so, several covariance matrices need to be evaluated. These are K, a matrix containing the covariances of all the possible combinations of $\mathbf{x}_1$ to $\mathbf{x}_n$, K_*, a vector containing the covariance of the training data with the new input point $\mathbf{x}^*$ and K_{**}, the covariance of the data at the new point $\mathbf{x}^*$:

$$K = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_1) & \cdots & k(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix}. \tag{4.6}$$

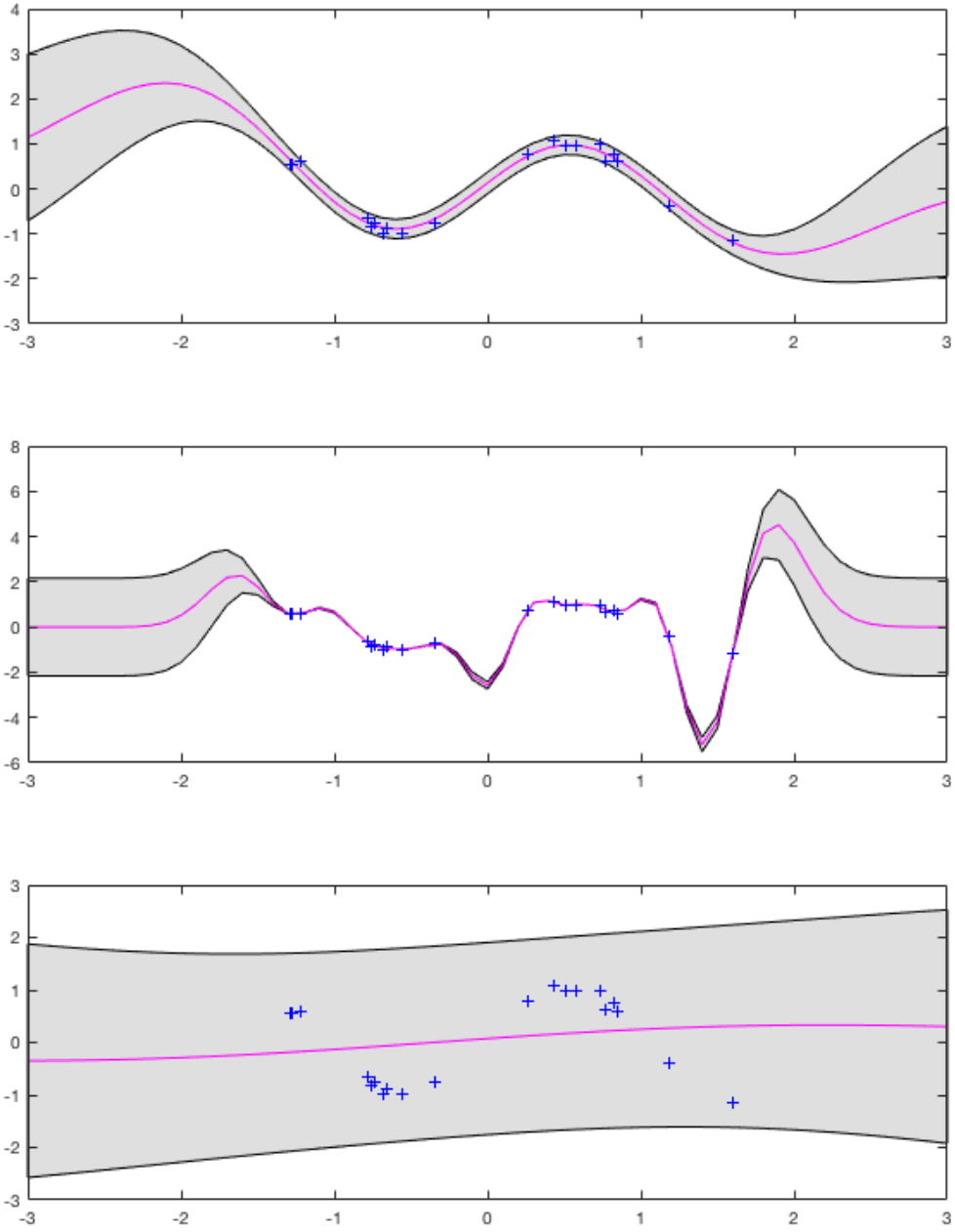


Figure 4.1: Picking the right hyperparameters (l, σ_f, σ_n) is crucial for Gaussian process regression.

Figure 4.1: *(previous page)*: Blue crosses are the training data, the underlying function fit to the data is the pink line, and the covariance used is the squared exponential function. The grey area corresponds to approximately two standard deviations from the mean for each value; the mean is 0. This 95% confidence area gets larger in the spaces where no training inputs exist to 'pin down' the function, such as the area after the last input and before the first input. The same training data is used for all 3 subplots. (a) The hyperparameters used are (1, 1, 0.1) - corresponding to the GP that generated the dataset. (b) The hyperparameters used are (0.3, 1.08, 0.001) - the shorter length scale means that the quickly varying function tends to pass through every point. The noise variance is very small here as the variation in the function is explained by the larger signal variance σ_f. (c) With hyperparameters (3.0, 1.16, 0.89), the long length scale describe a slowly changing function, with the noise variance increased in order to explain the deviations from the mean.

$$K_* = [k(\mathbf{x}^*, \mathbf{x}_1), k(\mathbf{x}^*, \mathbf{x}_2), ..., k(\mathbf{x}^*, \mathbf{x}_n)] \qquad K_{**} = k(\mathbf{x}^*, \mathbf{x}^*). \tag{4.7}$$

This is shown for the case where there is a single new input $\mathbf{x}^*$, but it can be extended for multiple new inputs $[\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_n^*]$ by extending K_* to a covariance matrix between the new input points and the training input points and K_{**} to a covariance matrix between the newly observed inputs.

From the definition of a GP, the training outputs $\mathbf{y}$ and predicted function values $\mathbf{y}_*$ have a joint multivariate Gaussian distribution. This is written as:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K + \sigma_n^2 \delta(\mathbf{x}, \mathbf{x}') & K_* \\ K_* & K_{**} \end{bmatrix}\right), \tag{4.8}$$

The probability function of $\mathbf{f}_*$, given the data already known, is normally distributed with mean and variance as follows:

$$\mathbf{f}_* | \mathbf{y} \sim \mathcal{N}(K_* K_\sigma^{-1} \mathbf{y}, K_{**} - K_* K_\sigma^{-1} K_*^\top), \qquad \text{where } K_\sigma = K + \sigma_n^2 \delta(\mathbf{x}, \mathbf{x}') \tag{4.9}$$

K^{-1} is the inverse of matrix K, $K_*^\top$ is the transpose of vector K_*, and our best estimate for function values of $\mathbf{f}_*$ is the mean, $K_* K_\sigma^{-1} \mathbf{y}$, corresponding to test inputs $\mathbf{x}^*$. These matrices, which were introduced in equations 4.6 and 4.7, can all be calculated and the values used to predict $\mathbf{f}_*$. This is simply done by sampling from a GP with mean function and covariance from equation 4.9.

In order to make predictions, it is therefore necessary to obtain estimates for the hyperparameters in order to be able to calculate the covariance matrices K, K_*, K_{**}. Point estimates of the hyperparameters can be obtained by maximising the marginal log likelihood: $\log p(\mathbf{y}|\mathbf{x}, \theta)$ where θ are the hyperparameters (Rasmussen, 2004). As the marginal log likelihood contains terms for complexity penalty and data fit, it automatically incorporates Occam's razor in the GP model. In this work, the marginal log

likelihood is maximised using a gradient descent optimisation tool: `http://learning.eng.cam.ac.uk/carl/code/minimize`.

4.3.2 Covariance kernels

Different kernels (or covariance functions) can be used to evaluate the covariance $k(\mathbf{x}, \mathbf{x}')$. The choice of kernel is important as it reflects prior information about the function connecting inputs and outputs. For instance, in modelling circadian genes, a periodic covariance function would be appropriate for capturing the sinusoidal rhythms of the genes. Choosing an appropriate kernel is a complex problem, especially when the underlying function is unknown.

The squared exponential (SE) function is popular as it makes a simple assumption that data points closer to each other are more highly correlated than data points further away from each other. The SE function is a specific case of the more general Matern kernel. The Matern kernel can specify the smoothness of the samples from extremely smooth (SE) to extremely rough (the Ornstein-Uhlenbeck process kernel). Another kernel, the rational quadratic, is an infinite sum of SE covariance functions with varying length scales. Other kernels include white noise (uncorrelated noise), constant (constant function), and linear (linear function) kernels.

In addition, kernels can be combined through addition and multiplication to model a richer set of structures such as data that are periodic and increasing over time (Lloyd *et al.*, 2014). Adding kernels is equivalent to a superposition of the independent functions. Multiplying kernels modifies their properties such as enforcing smoothness and periodicity to rough or non-periodic kernels. For instance, multiplying kernels by the SE locally smooths predictions from the original kernel and removes long range correlations. Change-points for functions that change suddenly can be formed by having kernels multiplied with sigmoidal functions, then added up. Additional change-point hyperparameters need to be specified which stipulate the position in time of the switch and the steepness of the switch. Change-point function are pertinent where the inputs experience a perturbation during the time period in which the data are captured. For instance, in infection data, the plant experiences two states: the normal uninfected baseline and the infected state where the plant defences are triggered by pathogen action.

4.3.3 The repressilator, a GPDS case study

A well known example of a gene regulatory network (GRN) in synthetic biology is the repressilator (Elowitz and Leibler, 2000), a synthetic three-gene system comprised of three mRNAs and three proteins as shown in Figure 4.2. The mutual repression of gene expression in this relatively simple network can give rise to complex nonlinear dynamics, which can be described by a coupled set of ODEs.

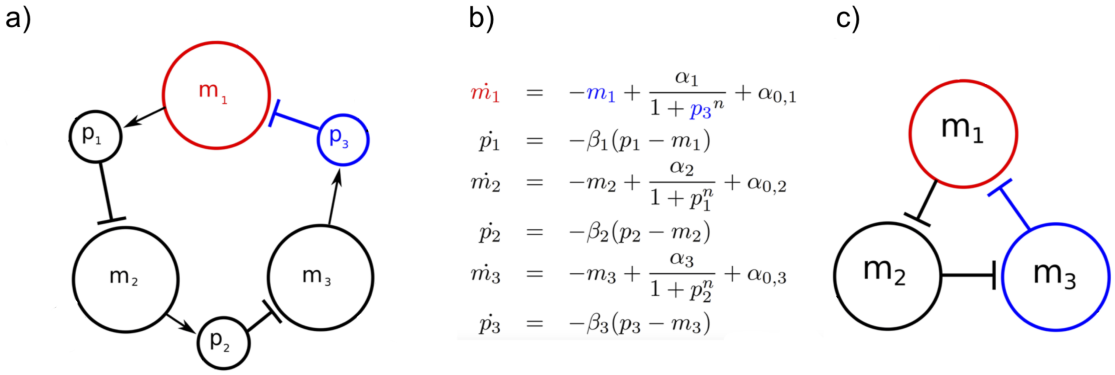


Figure 4.2: The repressilator synthetic three gene network. (a) Gene 1 encodes for mRNA 1 (m_1) that is translated into protein 1 (p_1), with gene 2 encoding for mRNA and protein 2 (m_2, p_2), and gene 3 encoding for mRNA 3 and protein 3 (m_3, p_3). Within the system, p_1 suppresses the expression of m_2, p_2 suppresses m_3, and p_3 represses the expression of m_1. This system of mutual repression can be represented as a six component network, with kinetics described by the system of ordinary differential equations (b). (c) As protein levels are not usually measured, the repressilator is often represented as a three component system of three genes that mutually repress one another; the system would therefore be defined by three observed molecular species $[m_1, m_2, m_3]$. Image from Penfold *et al.* (2019).

When the network structure is known, and all molecular species have been measured, a GPDS simply might simply replace the parameterised ODE model with a set of nonparametric approximations. For the repressilator system, for example, this is written as:

$$\begin{aligned} \dot{m}_1 &= f_1(m_1, p_3), \\ \dot{p}_1 &= f_2(p_1, m_1), \\ \dot{m}_2 &= f_3(m_2, p_1), \\ \dot{p}_2 &= f_4(p_2, m_2), \\ \dot{m}_3 &= f_5(m_3, p_2), \\ \dot{p}_3 &= f_6(p_3, m_3), \end{aligned} \tag{4.10}$$

where $\dot{m}_1$ is the rate $\frac{dm_1}{dt}$ and $f_i(\cdot)$ denotes an unknown, potentially nonlinear function, that describes the dynamics of molecular species i from the expression dynamics of its causal regulators. In this case, the unknown function, $f_i(\cdot)$, can be assigned a suitable prior distribution and a version of Bayes' rule is used to learn functions from data:

$$P(f_i|\mathcal{D}) = \frac{P(f_i)P(\mathcal{D}|f_i)}{P(\mathcal{D})},$$

where $\mathcal{D}$ is a dataset, $P(f_i|\mathcal{D})$ is the posterior distribution (that we aim to infer), $P(\mathcal{D}|f_i)$

is the marginal likelihood, $P(\mathcal{D})$ is a normalising constant, and $P(f_i)$ represents a prior distribution over functions. A suitable prior distribution over functions is given by the Gaussian process (GP).

In Figure 4.3, Gaussian process regression is illustrated using a simple example, for a one-dimensional function, $y = f(x)$. In Figure 4.3(a) the GP prior is indicated, showing the prior mean (solid black line) and 95% confidence intervals (CI) for y at various points of x. Samples can be drawn from the GP prior, shown in Figure 4.3(b). Given an observation, and using Bayes' rule, the GP prior can be updated to yield a posterior Gaussian process, performing nonlinear regression, as shown in Figure 4.3(c). Observations of y at x can be said to "pin down the GP". As the number of observations increases, the posterior GP begins to resemble the true underlying function, and the associated 95% confidence intervals become smaller; similarly, sample functions from the posterior GP begin to resemble the true function, as shown in Figure 4.3(d). In Figure 4.3(e) sample prior functions are shown for various values of the hyperparameter l. For a given hyperparameter and set of data, a posterior GP can be generated as shown in Figure 4.3(f). The marginal likelihood of the GP can also be evaluated, allowing the selection of the optimal hyperparameter values. In Figure 4.3(g) corresponding marginal likelihoods are shown for the posterior GPs in Figure 4.3(f). GPs can also be used to specify distributions over functions over higher dimensional spaces, and in Figure 4.3(h) an example posterior GP for a two dimensional function, $z = f(x, y)$, is shown, with the posterior mean of the GP indicated by the red surface, and an illustrative 95% confidence intervals shown as blue surfaces for a small region of the space.

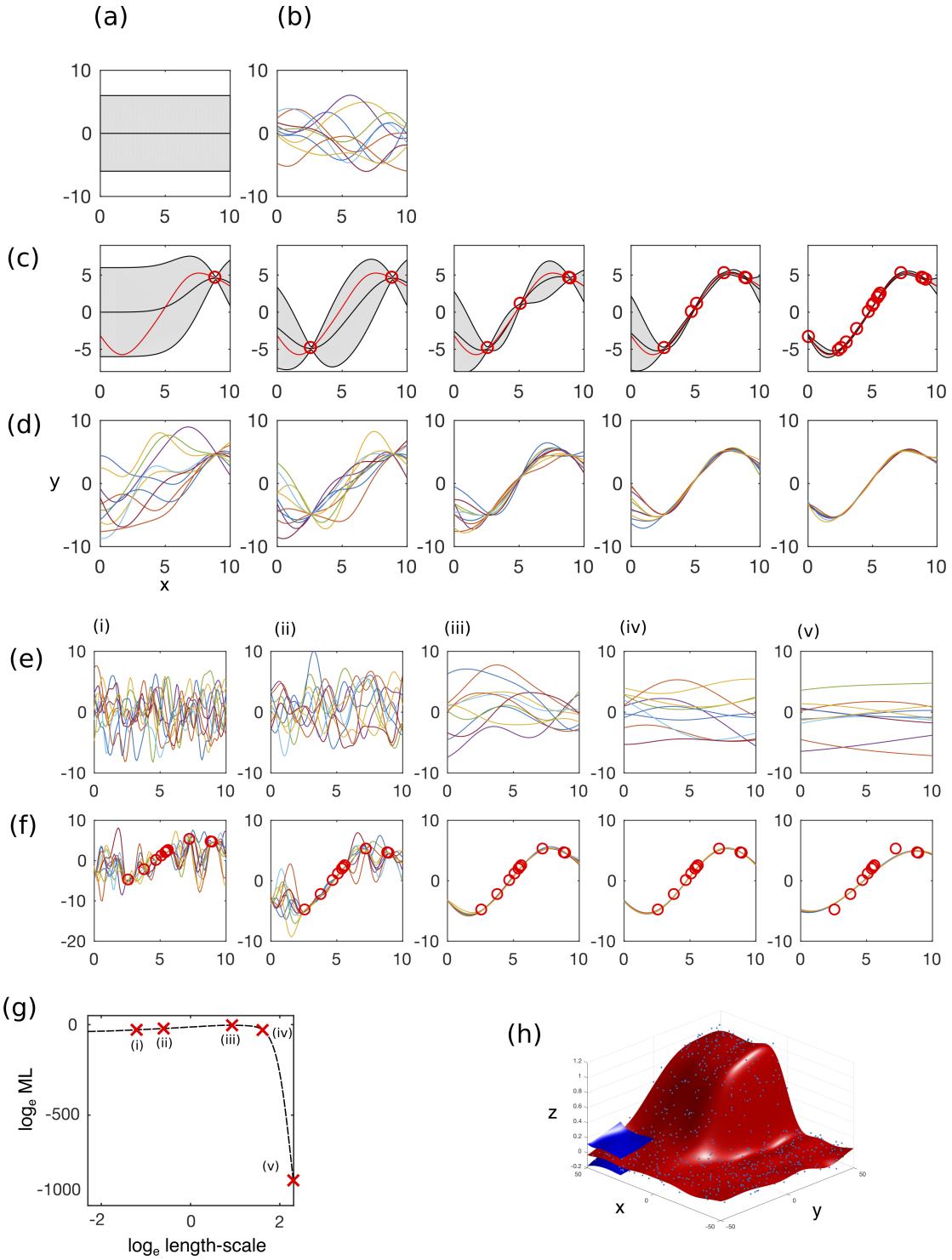


Figure 4.3: In a Bayesian setting, Gaussian processes (GPs) are infinite generalisations of the Gaussian distribution used to represent prior distributions over functions, allowing us to perform nonlinear regression. (a) The prior mean (solid black line) and 95% confidence intervals of f at various points of x. (b) Samples from the GP prior. (c) Nonlinear regression using a GP. As the number of observations are increased, the posterior GP begins to resemble the true underlying function, and the associated 95% confidence intervals become smaller.

Figure 4.3: *(previous page)*: (d) Similarly, sample functions from the posterior GP begin to resemble the true function. (e) Sample prior functions for various values of the hyperparameter l. (f) Posterior GPs for a given hyperparameter and set of data, (g) The corresponding marginal likelihoods for the posterior GPs indicated in panel (f). (h) An example posterior GP for a two dimensional function, $f(x, y)$, with the mean of the GP shown by the red surface, and illustrative 95% confidence intervals shown by blue surfaces for a small region of the space. Image from Penfold *et al.* (2019).

Since, in this particular example, the regulators are known, the (distribution over) the individual functions that govern each of the molecular species can be learned. For mRNA 1, for example, the regulators consist of m_1 and p_3 (Figure 4.2), and the aim is to understand how $\dot{\mathbf{m}}_1 = (\dot{m}_1(t_1), \ldots, \dot{m}_1(t_T))^\top$ is predicted by $\mathbf{m}_1 = (m_1(t_1), \ldots, m_1(t_T))^\top$ and $\mathbf{p}_3 = (p_3(t_1), \ldots, p_3(t_T))^\top$. This will enable the general prediction of $\dot{\mathbf{m}}_1^*(t)$ given new observations of $\mathbf{m}_1^*(t)$ and $\mathbf{p}_3^*(t)$, not previously seen by the model.To do so, the values of $\dot{\mathbf{m}}_1$ can be plotted against the corresponding $\mathbf{m}_1$ and $\mathbf{p}_3$ values in the phase space, as illustrated in Figure 4.4(a, b). Once the time series has been mapped into the appropriate phase space, inference of the function, $f_1(\cdot)$, becomes a regression task, and the (potentially) nonlinear function can be learned using Bayes' rule with a GP prior.

For a single time series, it seems unlikely there would be sufficient information to accurately capture the full range of dynamics available to m_1, and the posterior GP will therefore have large amount of uncertainty, as illustrated by the large 95% CIs in Figure 4.4(c). However, when there are multiple time series under a range of different initial conditions, these can be combined together. This extended set of observations can then be mapped on to the phase space (Figure 4.4(e)). As the number of perturbed time series increases, the data begins to span a much fuller range of the phase space, and the GP is able to capture the behaviour that allows generalisation to regions where no observations exist, with less uncertainty (Figure 4.4(f, g)).

4.4 Results

4.4.1 Gaussian processes capture DREAM network dynamics

The validity of Gaussian process regression for GRN modelling is first demonstrated by learning a DREAM4 Challenge (Marbach *et al.*, 2009b) 10-node and 100-node network and predicting gene expression for a validation dataset.

4.4.1.1 Simulating a DREAM10 wild-type network

In this chapter, a single DREAM network model of size 10 is used - DREAM10.1 (Figure 4.10). This un-rewired network is referred to as the wild-type (WT) network. The GPDS training data for learning hyperparameters and predicting new output values

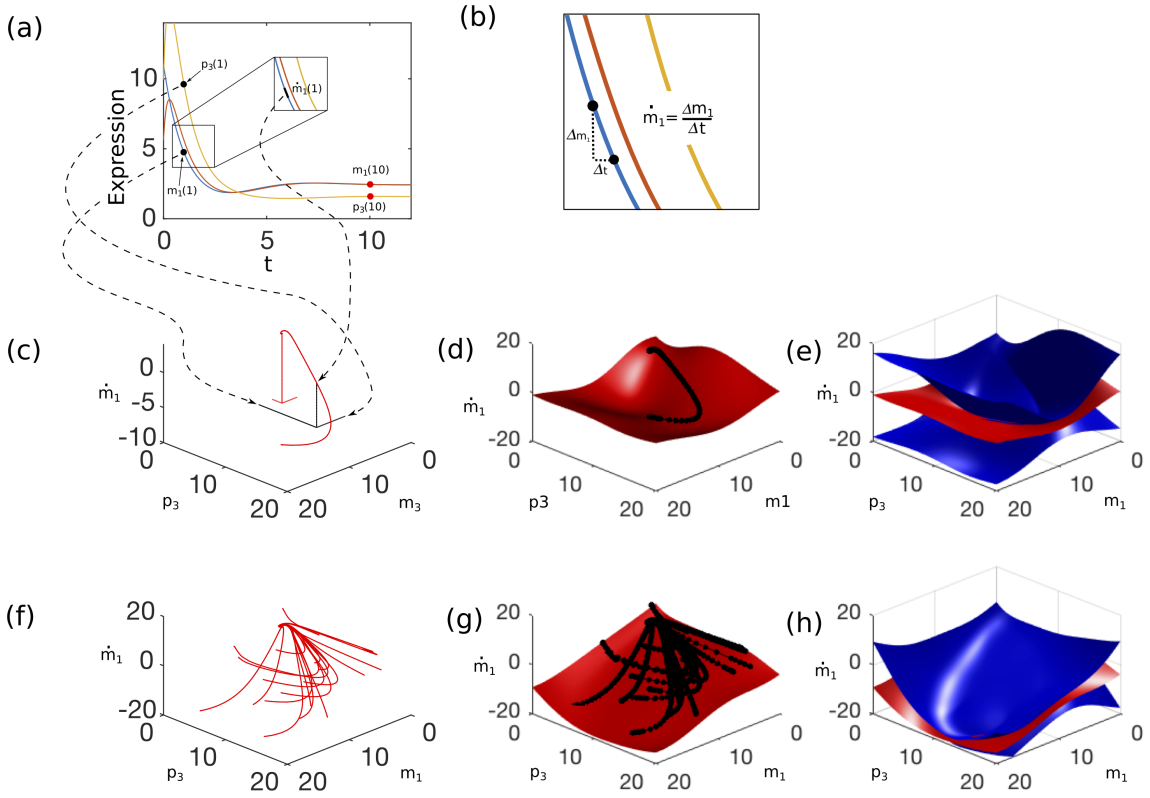


Figure 4.4: (a) Example expression profiles for mRNA m_1 (blue) and proteins p_1 (red) and p_3 (yellow) of the repressilator.

Figure 4.4: *(previous page)*: (b) Typically the derivative of the protein or mRNA expression levels cannot be measured directly, but instead must be calculated prior to inference (Äijö and Lähdesmäki (2009), Klemm *et al.* (2008),Penfold and Wild (2011)). (c) As the regulators of m_1 are known, the aim is to directly infer its dynamic behaviour from the data. This can be done by mapping the expression data into the phase space: by plotting $\dot{m}_1$ against m_1 and p_3. In (a, c) this mapping is shown explicitly for one data point at time point $t = 1$. Once the data has been mapped into the corresponding phase space, the behaviour of the system can be captured by fitting a Gaussian process model to the data. (d) Although a GP can be fit to a single time series data, as indicated by the mean function (red), the dataset is unlikely to contain sufficient information to be able to accurately predict behaviour at new locations, and the posterior GP has a large amount of uncertainty as indicated by the 95% CIs (e). (f) Multiple trajectoriescan be mapped, corresponding to different perturbed time series, into this phase space. (f, g) When there are enough observations, the GP can accurately capture the behaviour of the system, with the mean (g) closely approximating the true underlying function with reduced uncertainty (h). Image from Penfold *et al.* (2019).

are time series data from the DREAM network generated with GeneNetWeaver (GNW) Schaffter *et al.* (2011).

The training data available are 20 time series following 20 different perturbations on the network, with experimental noise added to generate 4 biological replicates - resulting in 80 time series of 21 time points each. The added measurement noise to create the biological repeats is similar to that found in microarrays (Marbach *et al.*, 2009a). These data are described in more detail in Section 2.3.1. The validation dataset is a single time series generated by using a different perturbation pattern to the training data. Unless otherwise stated, the value for the predicted outputs is the mean of $\mathbf{f}_*|y$ as described by equation 4.9. The fit of this mean GP prediction for new inputs is quantified by using a mean square error. This is:

$$\frac{1}{n}\sum_{i=1}^{n}(\sum(y_i - \hat{y}_i)^2), \tag{4.11}$$

where y_i are the predicted outputs of a certain gene, $\hat{y}_i$ the true outputs of that gene, and $n = 10$, the number of genes. The MSE represents the average variation between the GPDS simulated data and the GNW 'true' data for each gene.

Let X represent a gene of interest for which we wish to predict the expression value. The data from each time series are organised into training inputs - the gene expression values of gene X's regulators at time $t = 1$ to $t = 20$ - and training outputs - the expression values of gene X at time $t = 2$ to $t = 21$. It is assumed that the expression of the regulator at a certain time t influences the expression of its targets at the next time point $t + 1$.

At time point 1, the values of all the genes of the validation dataset are known,

and the challenge is to predict the values, one at a time, until t = 21. The output y^* is the value of gene X at time point $t + 1$. The inputs $\mathbf{x}^*$ are the values of the regulators of gene X at time point t. Current predictions are therefore used to inform future predictions. As mentioned in the introduction, different covariance kernels can be used to model the unknown function connecting inputs and outputs. 10 kernels were chosen for GPDS (the mathematical expressions for these are given in Appendix B):

- covSEiso - isotropic squared exponential covariance function
- covSEard - squared exponential covariance function with Automatic Relevance Determination (ARD)
- covRQiso - isotropic rational quadratic covariance function
- covRQard - rational quadratic covariance function with ARD
- covSEiso + covSEiso - composite covariance function made from the sum of two isotropic squared exponential covariance functions
- covSEiso * covSEiso - composite covariance function made from the product of two isotropic squared exponential covariance functions
- covSEiso + LIN - composite covariance function made from the sum of an isotropic squared exponential covariance function and a linear function
- covSEiso * LIN - composite covariance function made from the product of an isotropic squared exponential covariance function and a linear function
- CP Const + Const - change-point kernel using two constant covariance functions
- CP Const + covSEiso - change-point kernel using a constant and an isotropic squared exponential covariance function

Isotropic means the hyperparameters for each regulator regulating a target gene are the same, and Automatic Relevance Determination lets a function learn a separate hyperparameter for each regulator.

Given the training dataset and one of 10 covariance functions, the GPDS hyperparameter optimisation is carried out with a gradient descent optimisation tool that aims to maximise the marginal log likelihood (Rasmussen and Nickisch, 2010). Once the underlying distribution over the functions connecting regulators and targets has been learned, it is used to make predictions on the validation dataset. G1, G8 and G9, which do not have regulators in the original network, are not simulated - their values at each time point are provided in the regression problem. Their values are "fixed".

GPs were trained on different subsets of the training data, and then used to make predictions on the same validation dataset. The datasets had either 1 or 4 biological repeats (the same perturbation) and data from 1, 5, 10, or 20 different conditions (perturbations). The smallest MSE was achieved by using 20 time series representing 20

different conditions and a single biological repeat. The re-wiring described later on is carried out using the hyperparameters inferred from this particular training scheme.

Table 4.1: MSE of predictions made with GPDS for the DREAM10 WT network varies with the amount of training data provided.

Conditions	1	1	5	5	10	10	20	20
Biol reps	1	4	1	4	1	4	1	4
MSE	0.324	0.350	0.275	0.257	0.386	0.337	0.228	0.231

4.4.1.2 Ranking of covariance kernels

On the whole, the ten kernels performed comparably on the validation dataset; however the covRQard and covSEard functions performed particularly well. The is seen in the plots of the mean predictions for the validation dataset in Figure 4.5 and the MSE values in Figure 4.7. Meanwhile in Figure 4.5, 30 samples are drawn from the probability distribution $f_*|y$ (Equation 4.9). For some genes, such as G3, the uncertainty is higher and hence the samples are more spread out, while for others, such as G2, the variance is lower.

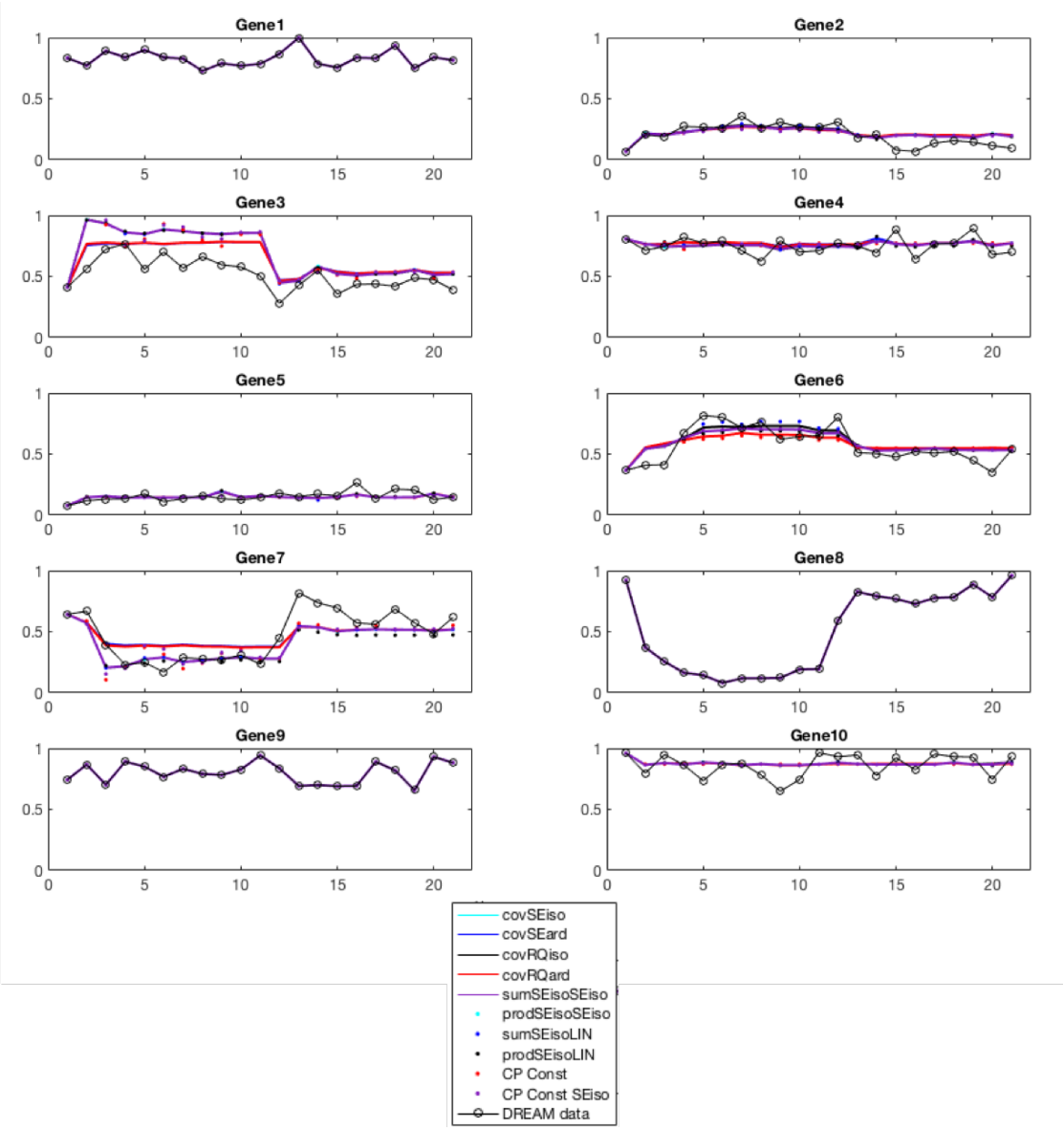


Figure 4.5: Performance of 10 covariance functions predicting the mean output y^* of the validation DREAM10 dataset. GPDS models learn the functions connecting regulators and targets from training data and predicted gene expression data for the validation dataset - which has the same underlying DREAM10 network but different perturbations applied to some of the genes. Gene expression data from the DREAM10 validation dataset is plotted in black with round markers, and simulations made using the 10 covariance functions defined above are also plotted in various colours. All models were trained on datasets with 20 different perturbations and 1 biological repeat. Simulations were started from the second time point onwards. Genes 1, 8 and 9 were not simulated.

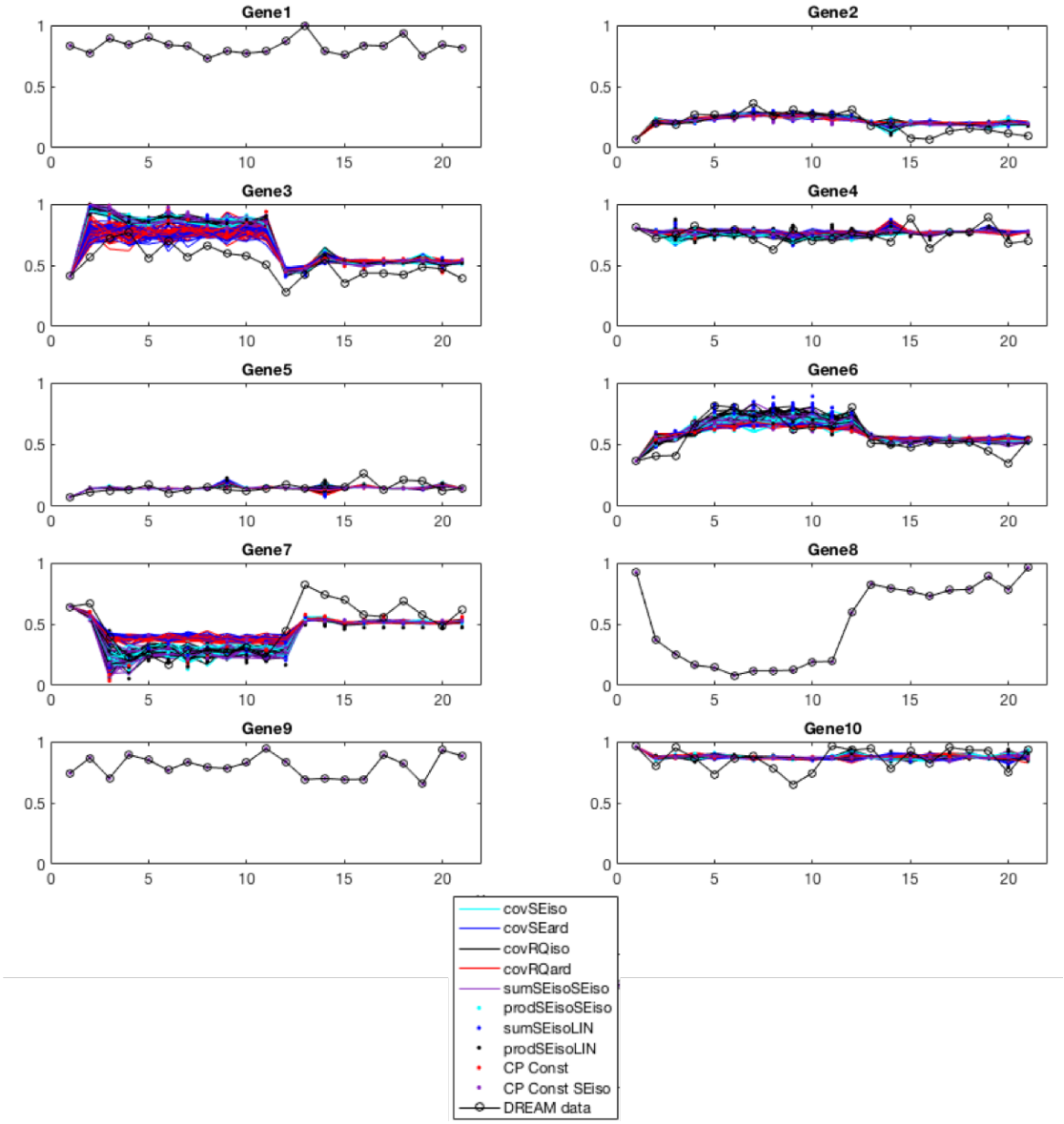


Figure 4.6: Performance of the GP model with 10 different kernels on the DREAM10 validation dataset. Gene expression data from the validation dataset is plotted in black with round markers, and simulations made using the covariance functions defined above are also plotted. The simulations are the same as in Figure 4.5, but here 30 samples are plotted from the GP posterior (with mean and covariance as in equation 4.9), whereas in Figure 4.5 only the mean of these samples is plotted. All models were trained on datasets with 20 different perturbations and 1 biological repeat. Simulations were started from the second time point onwards.

The MSE of each covariance in Figure 4.7 shows that the rational quadratic

kernel with ARD and the squared exponential kernel with ARD are best at predicting gene expression trends for this dataset.

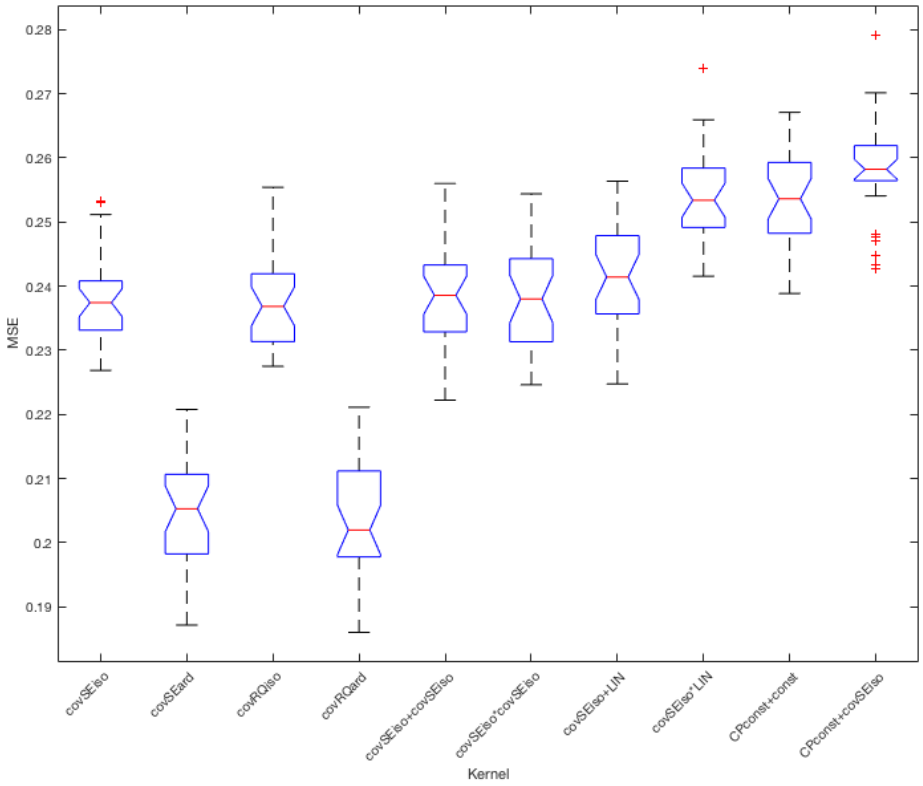


Figure 4.7: Boxplot of MSE for GPDS predictions, made with various kernels of the DREAM10 validation dataset. The MSE are calculated for each simulation shown in Figure 4.6 - 30 in all for each kernel. Kernel choice can have a significant difference on the accuracy of the resulting simulations, as determined by a 1-way ANOVA. The smallest MSE was achieved using the rational quadratic kernel with ARD, followed by the squared exponential kernel with ARD. In each box, the central red mark is the median, and the top and bottom blue edges are the 75th and 25th percentiles, respectively. Outliers are plotted individually using a '+'.

4.4.1.3 Simulating a DREAM100 wild-type network

The aim is to ultimately simulate large biological networks, therefore the performance of GPDS needs to be validated with a synthetic network larger than 10 nodes. The DREAM100 network of 100 nodes is used for this purpose. Similarly to the workflow for the DREAM10 network, the hyperparameters for the genes in the DREAM100.1 network are learned using gradient descent. The kernels used in this case are covSEiso, covSEard, covRQiso, covRQard, covSEiso + covSEiso, covSEiso * covSEiso, covSEiso +

LIN, and covSEiso * LIN. The training dataset consists of 10 different conditions with 21 time points each. The validation dataset consists of 21 measurements from a different condition to the training dataset (different pattern of perturbations applied to the random genes in the network). 14 of the 100 genes have no regulators and are therefore fixed. These are G5, G26, G36, G40, G42, G43, G46, G83, G90, G91, G93, G96, G98, G100.

Once again, the covRQard and covSEard kernels perform best as determined by the MSE. The resulting MSE on the validation dataset are shown in Figure 4.8. The first 20 gene simulations of the test data are shown in Figure 4.9, while the simulations for genes 21-100 are given in Appendix C.

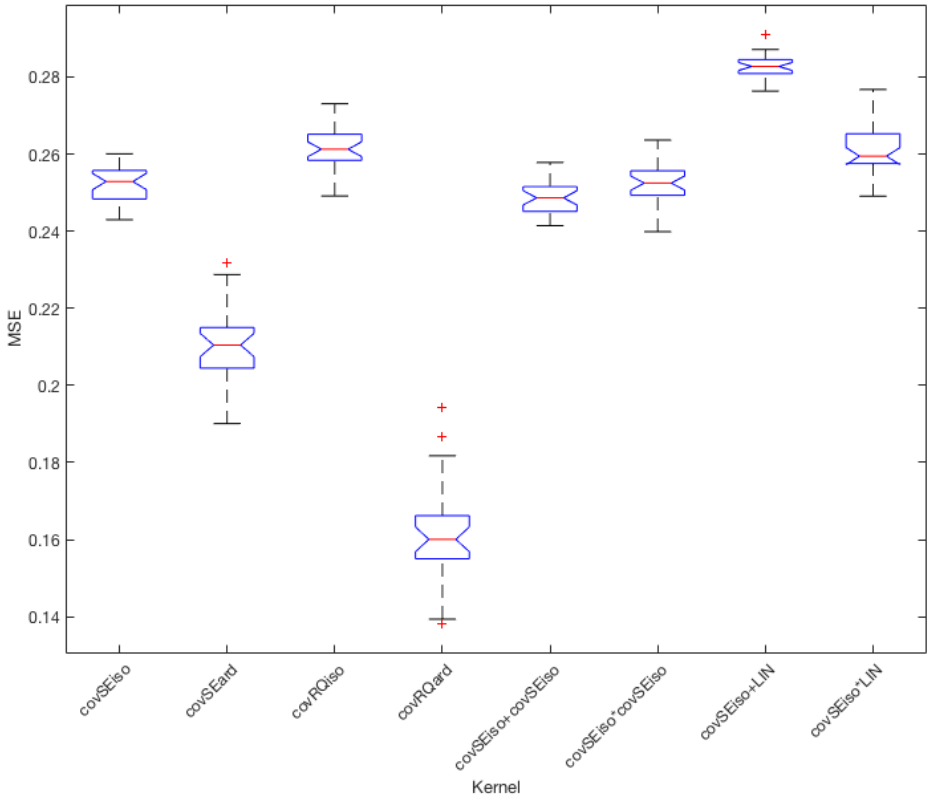


Figure 4.8: Boxplot of MSE for GPDS predictions, made with various kernels of the DREAM 100 validation dataset. The MSE are calculated for each simulation shown in Figure 4.9 - 30 in all for each kernel. Kernel choice can have a significant difference on the accuracy of the resulting simulation, as determined by a 1-way ANOVA. The smallest MSE was achieved using the rational quadratic kernel with ARD, followed by the squared exponential kernel with ARD. In each box, the central red mark is the median, and the top and bottom blue edges are the 75th and 25th percentiles, respectively. Outliers are plotted individually using a '+'.

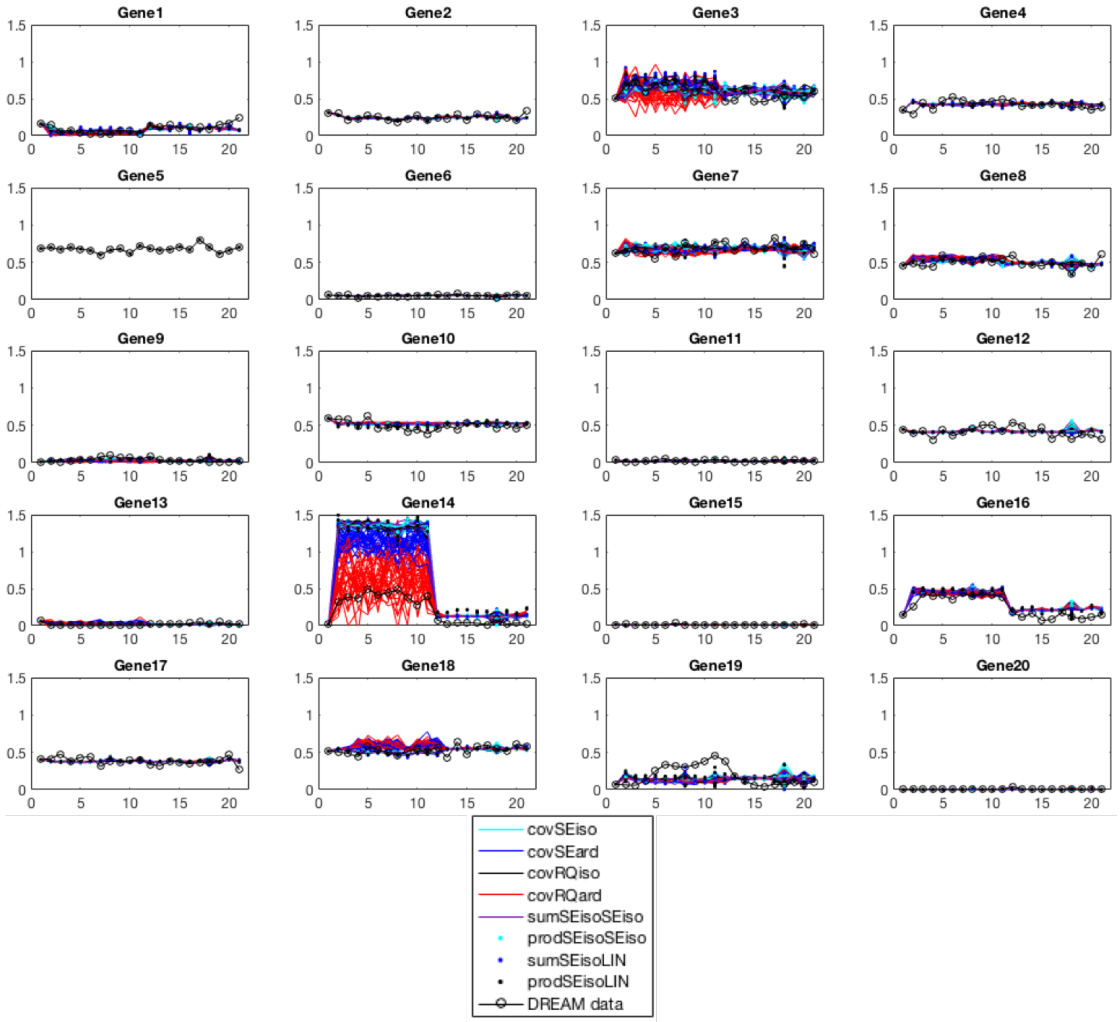


Figure 4.9: GPDS simulations using the 8 covariance functions for genes from the validation DREAM100 dataset. GPDS models learn the functions connecting regulators and targets from training data and made to predict gene expression data for the validation dataset - which has the same underlying network but different perturbations applied to some of the genes. Gene expression data from the validation dataset is plotted in black with round markers, and simulations made using the 8 covariance functions defined above are also plotted. All models were trained on datasets with 10 different perturbations and 1 biological repeat. Simulations were started from the second time point onwards. Only simulations for the first 20 genes out of 100 are shown here. G5 (among others not plotted here) is fixed so its levels are not simulated - its actual value is seen by the algorithm.

4.4.2 Modelling re-wired DREAM networks with Gaussian process dynamical systems

Subsequently, GPDS models were used to predict the gene expression levels in the DREAM10 and DREAM100 networks upon rewiring, with training data solely from WT networks. The GPDS models learned in the previous section were used for this purpose. DREAM networks are excellent for testing out inference and prediction methods as the gold standard true network is known and true gene expression values are obtained by using GeneNetWeaver.

4.4.2.1 Re-wiring of the DREAM10 network

In order to simulate rewirings, the top performing kernels were used in the GPDS: covSEard, covRQard, along with covSEiso (as it is the most popular standard kernel).

GeneNetWeaver was used to simulate time series with identical perturbations to the original DREAM data but where the network has been changed slightly, or rewired. The rewiring is done by swapping all the regulatory edges of a node with the regulatory edges of another node. For example, the regulators of G2 are {G1,G6,G8} and the regulators of G5 are {G1}. One possible rewired network is created by replacing the regulators of G5 with the regulators of G2 (see Figures 4.10 and 4.11). For the sake of brevity, the gene that is being rewired is referred to as the rewiring target, and the gene whose regulators are used for the rewiring (and is itself unchanged) is referred to as the rewiring source. In this example, G5 is the rewiring target and G2 is the rewiring source. The hyperparameters for G5 are then copied from G2. The hyperparameters represent how regulators affect the expression of the target gene and are therefore associated with a specific promoter region. If the promoter region of G5 is replaced with that of G2, therefore the relationship between G5 and its regulators should mimic the relationship between G2 and its regulators.

The model is trained using data from the wild-type DREAM10.1 network. The training data for G5 is a combination of the original training inputs and outputs for G5 and the training inputs and outputs for G2. The training data, hyperparameters and regulators for all the other genes remain the same. When they are not the target of a rewiring, the true rewired expression profiles of genes G1, G8 and G9 are assumed to be known, or are 'fixed'. GPDS are then used to predict the effect of this rewiring and compare it to the GNW 'true' values.

All possible rewirings are modelled by replacing the regulators of G1 to G10 with the regulators of G2, G3, G4, G5, G6, G7, or G10, minus redundant rewirings, such as replacing the regulators of G2 with the regulators of G2. The MSE is used as an indicator of the accuracy of the simulations. This is calculated by subtracting the simulated values from the GNW values, and squaring this residual. These are then summed, and divided by the number of genes being simulated (see Equation 4.11).

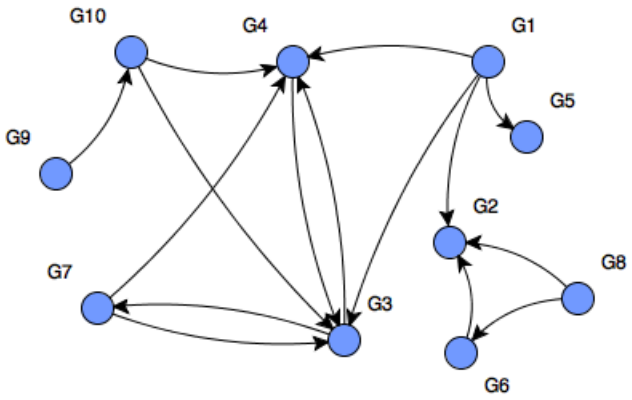


Figure 4.10: The network DREAM10.1 from the DREAM4 network challenge. The regulators of G2 are {G1,G6,G8} and the regulator of G5 is {G1}.

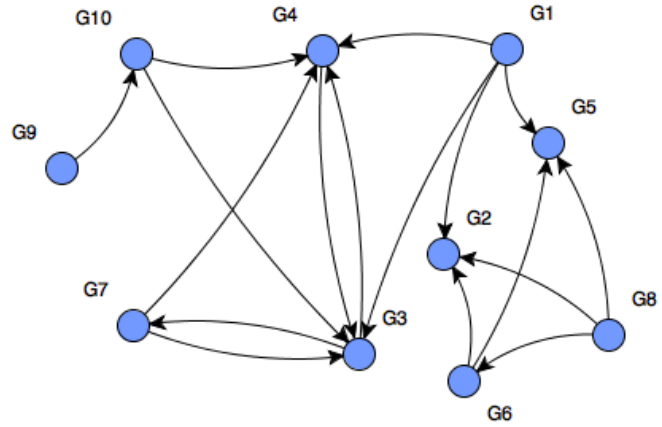


Figure 4.11: A rewired version of the network shown in Figure 4.10. The regulators for G5 are replaced with the regulators for G2.

The average MSE for the covSEiso, covSEard and covRQard kernels is 0.924, 0.916 and 0.914 over 315 simulated rewirings, respectively. However, 75% of the rewirings have a MSE smaller than 1. In Figures 4.12, 4.13 and 4.14, a typically 'good' (MSE of $\sim$ 0.5), 'average' (MSE of $\sim$ 0.9), and 'bad' (MSE of $\sim$ 4 - the largest MSE of all rewirings) performance are shown.

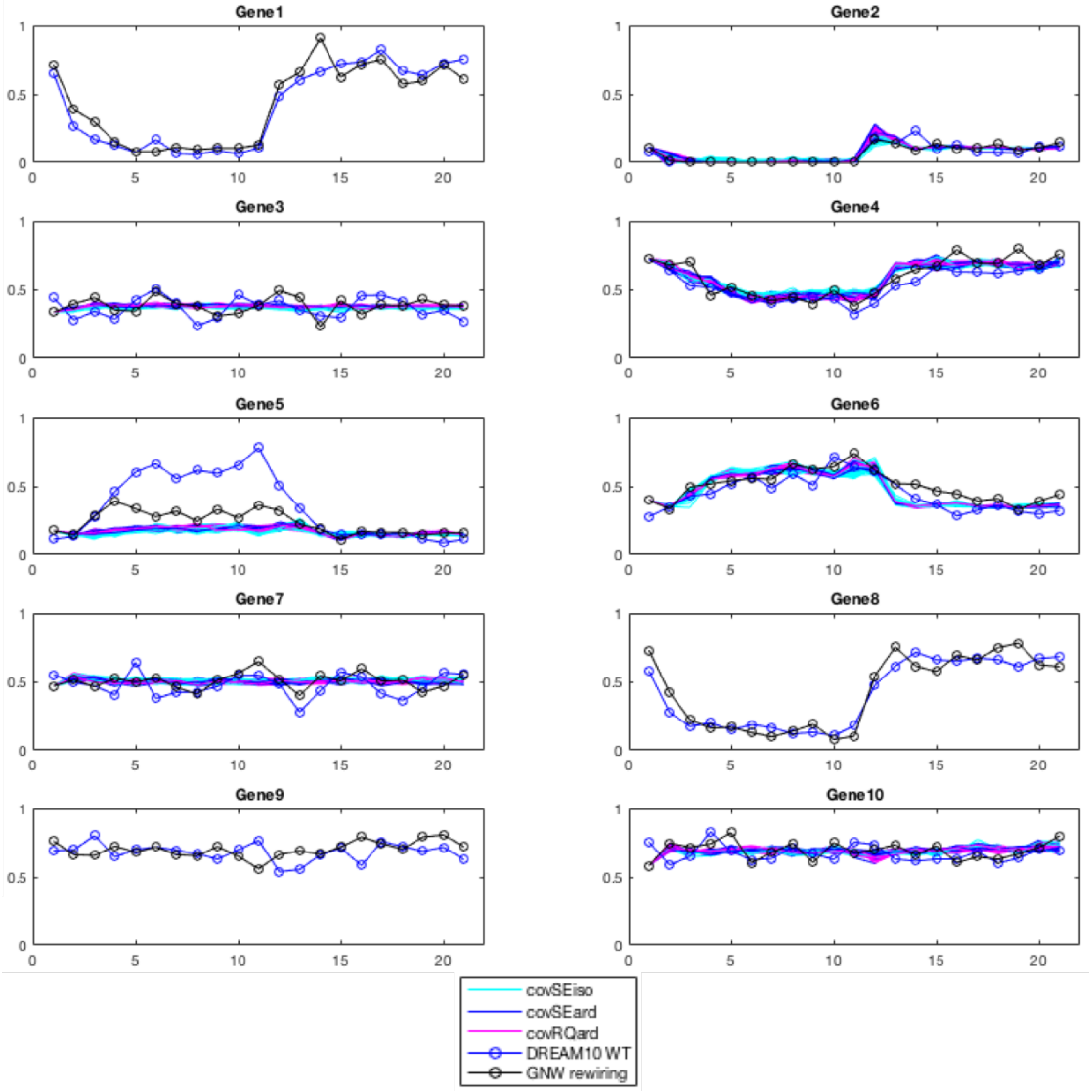


Figure 4.12: GPDS simulations for the rewiring of G5 with the regulators of G2 in the DREAM10 network, for the third perturbation. 30 samples are plotted from the GP posterior (with mean and covariance given in 4.9) for each kernel. Solid black line with circular markers: the gene expression data of the rewired network as given by GNW. Solid blue line with circular markers: the gene expression of data from the WT network with the same perturbation as the rewired network. Cyan, blue and pink solid lines are simulations from the covSEiso, covSEard and covRQard covariance kernels for the rewired network. Simulations are started off at the second time point and G1, G8 and G9 are fixed.

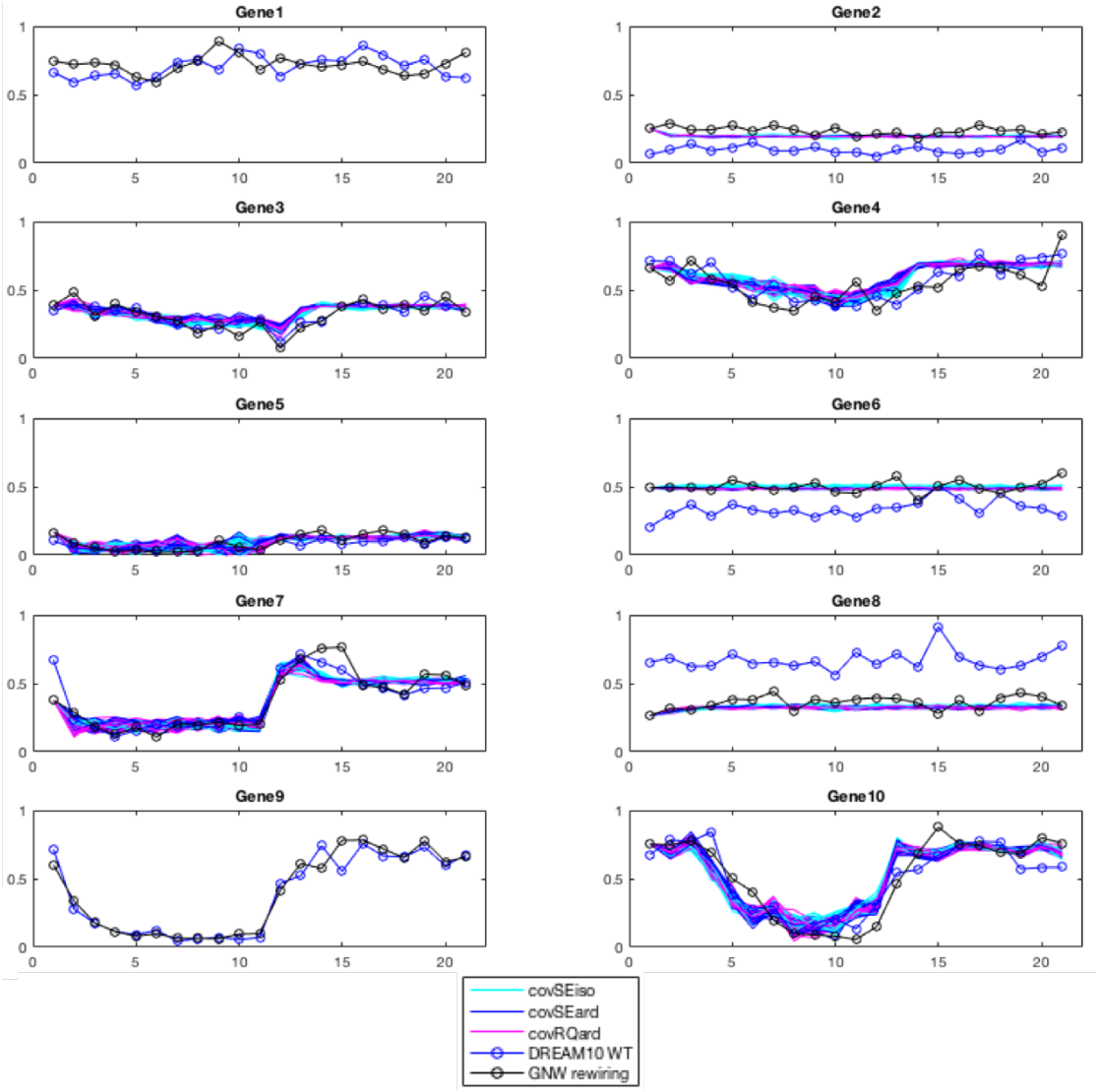


Figure 4.13: GPDS simulations for the rewiring of G8 with the regulators of G2 in the DREAM10 network, for the fifth perturbation. 30 samples are plotted from the GP posterior (with mean and covariance given in 4.9) for each kernel. Solid black line with circular markers: the gene expression data of the rewired network as given by GNW. Solid blue line with circular markers: the gene expression of data from the WT network with the same perturbation as the rewired network. Cyan, blue and pink solid lines are simulations from the covSEiso, covSEard and covRQard covariance kernels for the rewired network. Simulations are started off at the second time point and G1 and G9 are fixed.

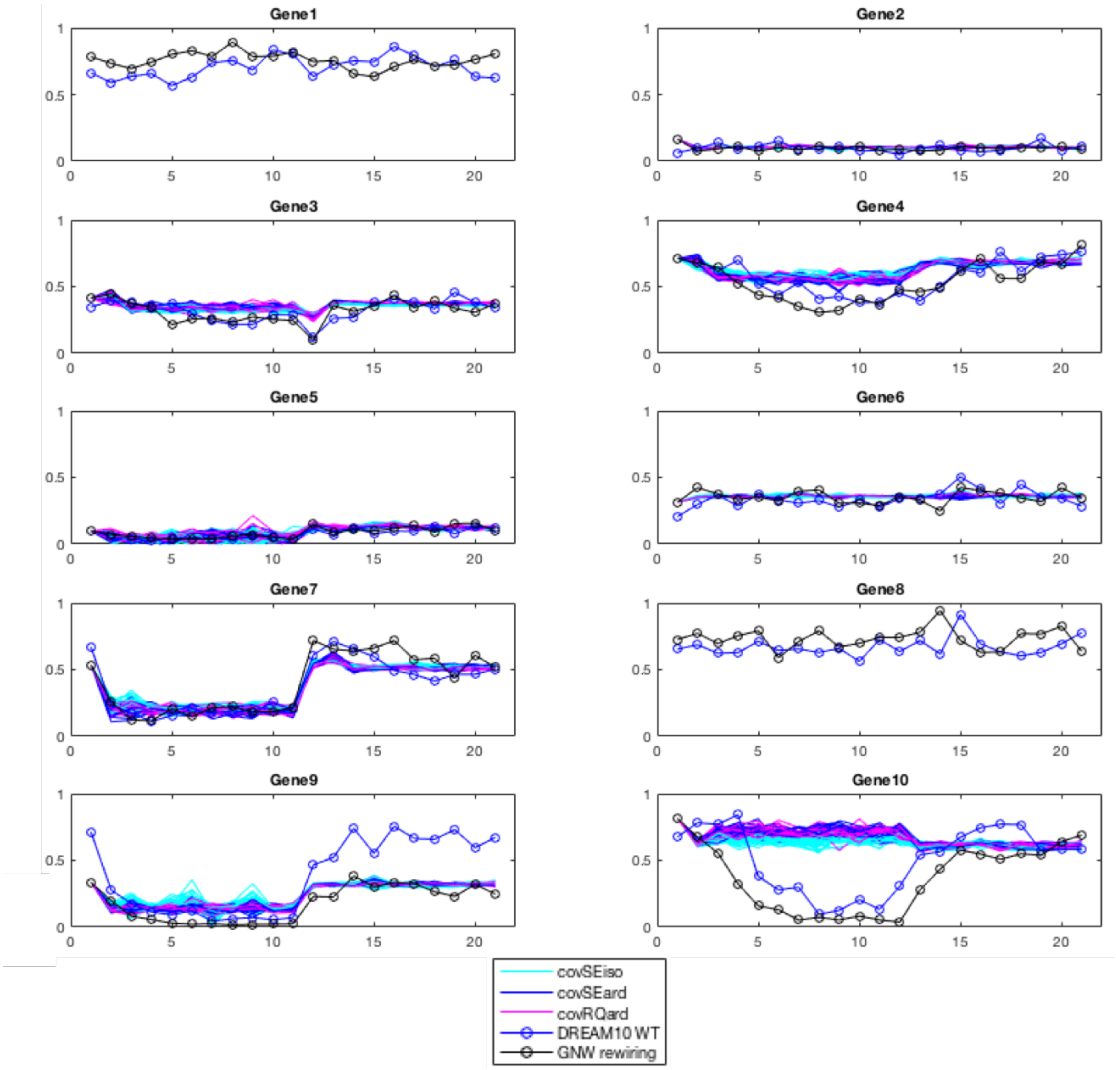


Figure 4.14: GPDS simulations for the rewiring of G9 with the regulators of G6 in the DREAM10 network, for the fifth perturbation. 30 samples are plotted from the GP posterior (with mean and covariance given in 4.9) for each kernel. Solid black line with circular markers: the gene expression data of the rewired network as given by GNW. Solid blue line with circular markers: the gene expression of data from the WT network with the same perturbation as the rewired network. Cyan, blue and pink solid lines are simulations from the covSEiso, covSEard and covRQard covariance kernels for the rewired network. Simulations are started off at the second time point and G1 and G8 are fixed.

Some rewirings have very little affect on the network other than the node that is getting rewired. For instance, rewiring G5 with G2 (Figure 4.12) only affects the levels of G5. However, rewiring G8 with G2 also affects G2 and G6 as they are directly

downstream of G8. If this happens and the levels of the downstream genes are fairly constant, the GPDS is able to reproduce that (Figure 4.13). However, in some cases, the GPDS has not accurately captured the relationship between the regulators and its targets (G10 in Figure 4.14).

Having more fixed genes improves the accuracy of the simulations. For instance, when G1, G8 or G9 are being rewired (so there are only 2 fixed genes), the average MSE for the covSEiso, covSEard and covRQard kernels is higher at 1.2476, 1.2425 and 1.2411, respectively. Otherwise, the average MSE for all rewirings is 0.9240, 0.9157 and 0.9140 for the covSEiso, covSEard and covRQard kernels, respectively. A closer look at the 33 individual rewirings that have a MSE of more than 1.5 reveals some patterns. For instance, in the rewiring of G1 with the regulators of G3 or G5 or G7, G4 and G5 do not increase or decrease as expected, but their levels stay flat (Figure 4.15 - top). Interestingly, a number of rewirings seem accurate at first glance, such as the rewiring from G2 to G7, which results in the levels of G7 decreasing. Indeed, this follows the expression pattern of G2 - however GNW predicts that the levels of G5 should remain high (Figure 4.15 - bottom). Such discrepancy is probably due to the difference in degradation rates of the two mRNA species. In the original ODE equation, the degradation rate of G2 is higher than that of G7, therefore, after rewiring, its levels do remain high. However, this subtlety was not captured by the GP function.

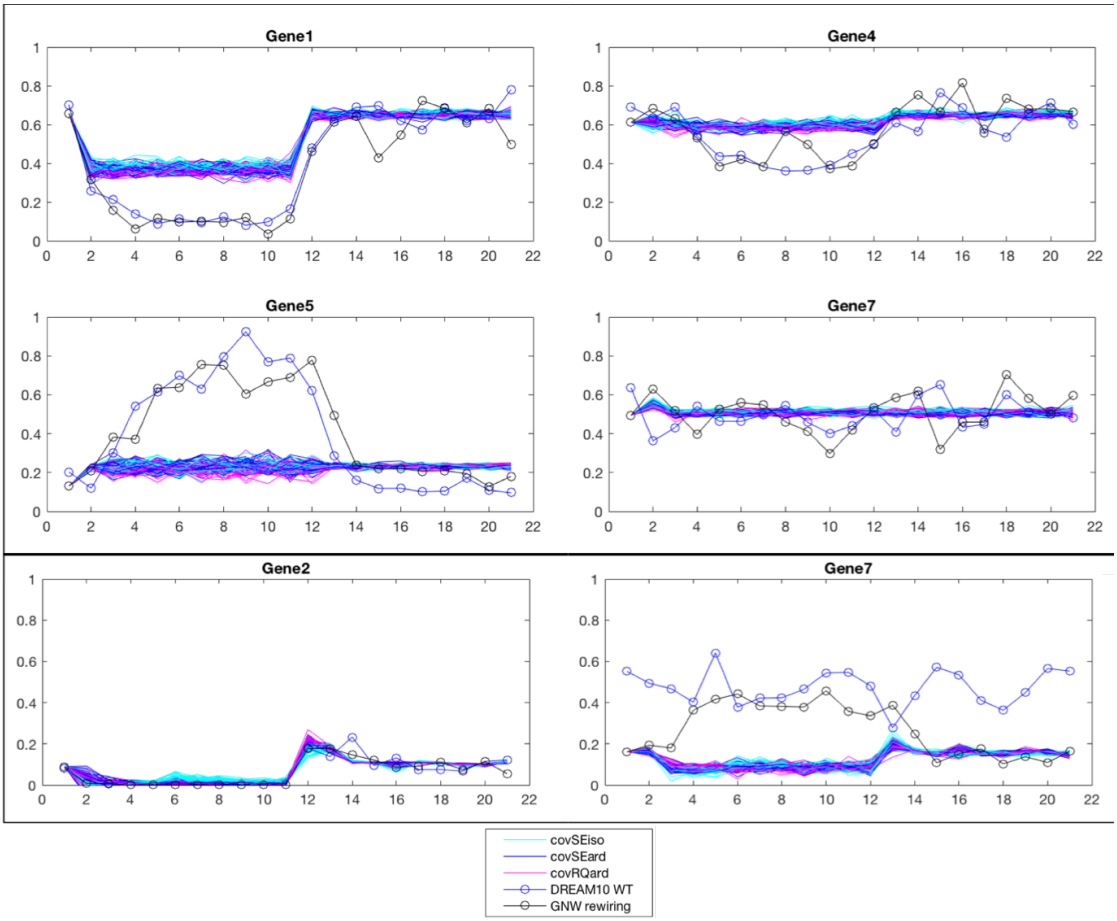


Figure 4.15: GPDS simulations for the rewiring of G7 with the regulators of G1 in the DREAM10 network, for the first perturbation (top box). GPDS simulations for the rewiring of G5 with the regulators of G4 in the DREAM10 network, for the first perturbation (bottom box). 30 samples are plotted from the GP posterior (with mean and covariance given in 4.9) for each kernel. Solid black line with circular markers: the gene expression data of the rewired network as given by GNW. Solid blue line with circular markers: the gene expression of data from the WT network with the same perturbation as the rewired network. Cyan, blue and pink solid lines are simulations from the covSEiso, covSEard and covRQard covariance kernels for the rewired network. G4 and G5 do not show the expected dynamic profile in the rewiring of G7 with G1 (top). While G7 follows the profile of G2 after being rewired using the regulators of G2, that is not the expected outcome as modelled by GNW (bottom).

4.4.2.2 Gene expression optimisation for the DREAM10 network

Once all possible rewirings for the DREAM network are simulated, it is useful to determine which rewiring results in network optimisation - the increase in level of certain

genes and a decrease in level of other genes, while keeping the rest of the genes in the network constant. In the *At-Botrytis* network, we are ultimately looking to optimise the expression of genes in the latter half of the time series, when infection is occurring. Similarly, in the DREAM network, the optimisation criteria is applied to the second half of the time series. If the average of the gene expression of Gene X at time points 12 to 21 in the rewired network is higher than its gene expression at those same time points in the wild-type network, the rewiring can be said to have increased the levels of Gene X:

$$f(x) = \begin{cases} \text{increase} & \text{if } \left(\frac{1}{10} \sum_{t=12}^{t=21} Z_t^x - Y_t^x\right) > 0.1 \\ \text{decrease} & \text{if } \left(\frac{1}{10} \sum_{t=12}^{t=21} Z_t^x - Y_t^x\right) < 0.1 \end{cases} \tag{4.12}$$

where Z_t^x is the simulated expression of the rewired gene x at time t and Y_t^x is the simulated expression of gene x at time t in the WT network. If the change in the level of the gene is smaller than 0.1 (10% of the full expression range), then it is not counted as differentially expressed.

When optimising Arabidopsis networks, we want the levels of known defence genes to go up and the levels of genes that have a detrimental effect on the plant to go down. The other genes of unknown function should be kept as close to their original levels as possible. Within the DREAM network, the genes do not have a particular phenotype so a few genes are selected at random for the optimisation.

The covRQard rewiring simulations are used in this case as they are the most accurate as quantified by the MSE. The task is to identify the rewirings that fit the desired criteria for all 10 genes, using the simulated rewiring values. The true best rewirings can be determined from the GNW rewired data. For instance, the optimisation function correctly identifies the 2 rewirings out of the total 315 that achieve the following criteria:

- Decrease G1,
- Increase G5,
- Maintain the values of the remaining genes

These rewirings are replacing regulators of G1 with regulators of G5 or G6. Similarly, the rewirings that decrease both G4 and G10 and leave the rest unchanged are correctly identified as rewiring the promoters of G5 and G6 to G10.

In some cases, it is not possible to achieve the desired optimisation while maintaining the WT levels of most of the genes. In this case it is still desirable to identify the rewirings that fit most of the criteria, while giving preference to the network configurations that fine-tune the levels of genes with a phenotype rather than those that maintain WT levels of all the genes. For instance, in the first example above, rewirings that increase the levels of G5 and consequently increase the levels of G10 are better than rewirings that keep the levels of G10 and G5 unchanged (all else being equal).

	TPR 9 criteria	TPR 8 criteria
Inc G3, dec G6	-	0.800
Inc G7, dec G10	0.970	0.920
Inc G2, dec G9	0.800	0.800
Dec G4, dec G10	1.000	0.790
Inc G3, inc G5	-	0.824

Table 4.2: True positive ratio (true positives/(true positive + false positives)) for the optimisation procedure identifying rewirings that fit 8 out of 10 (TPR 8 riteria) and 9 out of 10 criteria (TPR 9 criteria). 5 different optimisation schemes are given here. The genes not mentioned in the criteria are to be kept the same as WT levels. The first and fifth scheme do not have any rewirings that fulfill 9 out of 10 of the criteria. Inc increase; dec decrease.

4.4.2.3 Re-wiring and optimisation of the DREAM100 network

While the DREAM10 network is useful for a proof-of-concept that GPDS can be used to accurately predict the effect on rewirings, it is small compared to real-life GRNs. The DREAM100 network offers a better perspective on what to expect from rewiring a larger network, such as the ones generated in this thesis in Chapter 3.

In a similar fashion to the rewiring of DREAM10, the GPDS model is trained on the input and output data of the WT DREAM100 model. The target gene of the rewiring is also given the input and output data of the rewiring source gene as part of the training. Similarly, GNW is used to simulate what the true gene expression values are upon rewiring. The kernel used for simulating DREAM100 rewirings is the covRQard.

Not all possible rewirings are simulated, as that would be too costly in computational terms. Instead, a subset of genes are selected for rewiring: those that regulate more than 3 genes, so that the rewiring can have a large impact on the network. This is a total of 11 genes: G5, G26, G37, G44, G46, G63, G72, G83, G85, G93, G96. 4 genes are selected as the rewiring source: G1, G4, G24, G28. This leads to a total of 44 rewirings being simulated. The 14 genes without regulators in the WT network are fixed. The result of one such rewiring is shown in Figure 4.16 for the first 20 genes (simulations of genes 21-100 can be found in Appendix C). The average MSE for all the genes in all the rewirings is 0.0984.

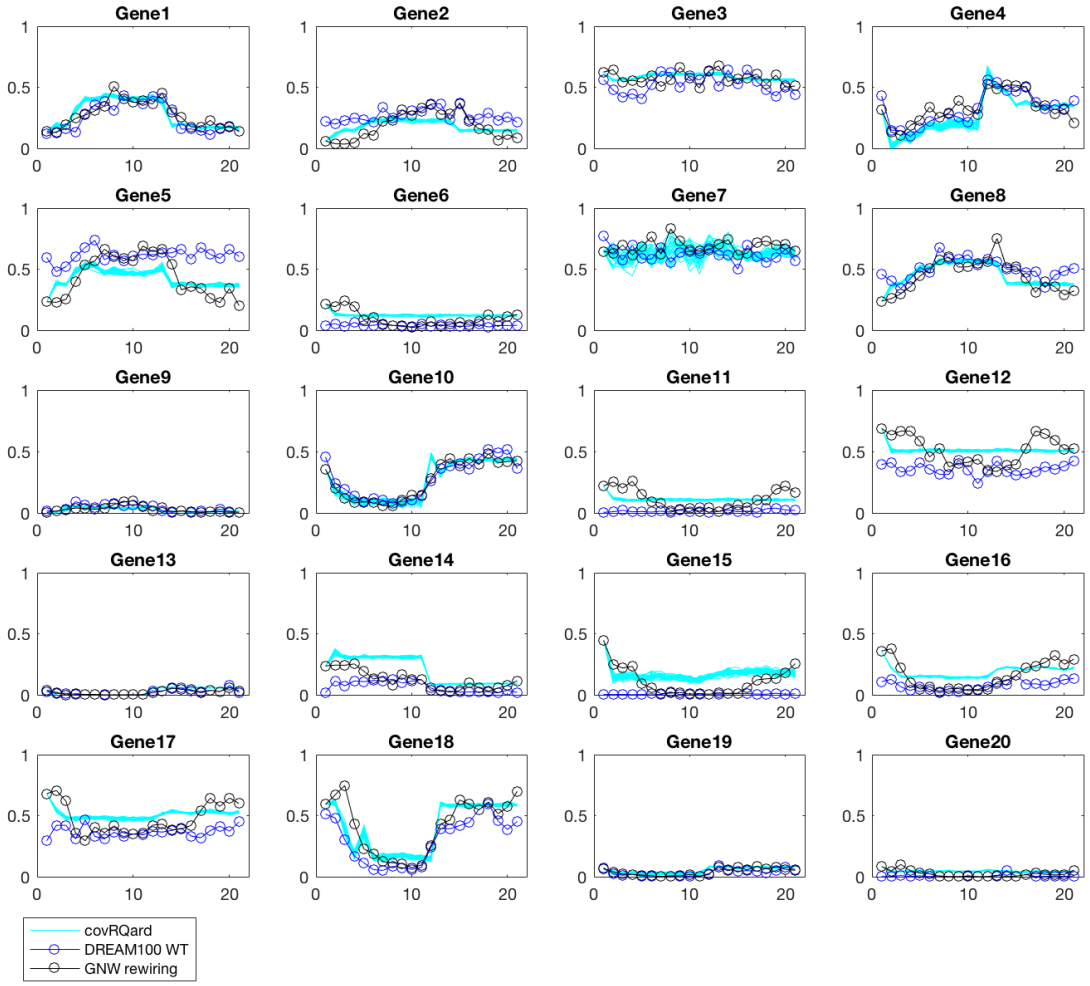


Figure 4.16: GPDS rewiring simulations of DREAM100 network gene expressions obtained by replacing regulators of G5 with the regulators of G1. Solid black line with circular markers: the gene expression data of the rewired network as given by GNW. Solid blue line with circular markers: the gene expression of data from the WT network with the same perturbation as the rewired network. Cyan solid lines are 30 simulations from the covRQard covariance kernel for each gene. Simulations are started off at the second time point and some genes in the DREAM100 network are fixed as outlined above.

Rewiring the DREAM100 network results in fewer differentially expressed genes (DEG) than rewiring the DREAM10 network - 8% of the DREAM100 genes are DE on average, whereas 17% of DREAM10 genes are DE on average. This is probably due to the increased robustness of the larger network. The differential expression of genes after rewiring is calculated via equation 4.12, with one change: the average is calculated over the whole gene expression, rather than just the latter half.

As demonstrated in Section 4.4.2.2, an optimisation function is applied to find the rewirings that have the desired effect on all or most of the 100 genes. Given the criteria to decrease G5, and increase G12, G15, and G17, the 10 rewirings that have a score of 93/100 and above are correctly identified using the GPDS simulated data - with 4 false positives, giving a precision of 1 and a true positive ratio of 0.7143.

4.4.3 Predicting rewirings to improve stress tolerance in Arabidopsis

4.4.3.1 Constructing the Arabidopsis 70GRN

The main aim of this project is to rewire Arabidopsis stress networks to enhance the plant's fitness during *Botrytis cinerea* infection. The proof-of-concept rewiring was carried out on the DREAM10 and 100 networks, and shown to be accurate enough to attempt rewiring the *At-Botrytis* consensus network inferred in Chapter 3.

In order to model the *At-Botrytis* network, it first needs to be reduced to a smaller number of nodes. However, unlike with the ODE simulations, it is not necessary to be as stringent with the network size.

The way this network reduction has been achieved is based on RNAseq data presented in Chapter 6. The simulations are to be used to narrow down all the possible rewirings to some promising rewiring candidates. The rewiring is to first take place in protoplasts for further screening. Using protoplasts for a medium throughput screening of rewiring constructs saves time in generating stable Arabidopsis transformants, which may take 6 months or more to reach T3 - the third generation of transformants which are homozygous for the genetic mutation. On the other hand, extracting and transforming protoplasts with plasmids takes two days, and the results of rewiring can be assessed either through reporter genes or qPCR. The benefits of using protoplasts are further discussed in Chapter 6. In order to mimic *B. cinerea* infection, transformed and control protoplasts can be treated with chitin for 45 minutes. Chitin is a component of the cell wall in fungi and elicits a defence response from Arabidopsis. As a microbial-associated molecular pattern, chitin triggers pattern-triggered immunity (PTI), the core basal immune response, and is recognised by surface-localised pattern-recognition receptors (PRRs) (Newman *et al.*, 2013).

However, not all genes behave in the same way in protoplasts that have been treated with chitin as Arabidopsis leaves that have been infected with *B. cinerea*. Pathogen infection triggers differential expression in many more genes (Windram *et al.*, 2012) than chitin treatment does (Libault *et al.*, 2007; Povero *et al.*, 2011; Ramonell *et al.*, 2005). Additionally, the process of digesting the plant cell wall with enzymes to release protoplasts is stressful and results in transcriptional reprogramming (Section 5.4.2) which can mask the transcriptional response to an additional stimulus, chitin. For instance, the internalisation of the FLS2 receptor occurs in protoplasts regardless of the presence of flg22 (Ali *et al.*, 2007). This priming of the defence response may reduce the number of DEG before and after chitin treatment.

In order to build models using data from *B. cinerea* infection of leaves, and make predictions to be tested out in protoplasts treated with chitin, it is important to limit

the genes in the models to those that act in the same manner in both systems. That is, if a gene is differentially expressed in the same way (increased or decreased) in both conditions, then it is retained in the model. This may eliminate regulators which play an important role in the TFs remaining in the model.

The details of how the DEG in Arabidopsis protoplasts following chitin induction for 45 minutes are determined are further explained in the Materials and Methods section of Chapter 5. A cut-off of adjusted p-value < 0.06 was used to selecting DEG from the protoplast experiment.This list of DEG was then compared to the list of DEG transcription factors (TFs) from the *B. cinerea* experiment (Windram *et al.*, 2012) and only genes that were present in both were selected. The less stringent cut-off was chosen as it was more important to avoid false negatives than false positives - particularly as the overlap between the two datasets was small. All the time points were included from the leaf time series again in order to maximise the overlap between protoplast and leaf DEGs. However, chitin only triggers the PTI, whether the ETI is also seen in a leaf infected with *B. cinerea*, particularly at the later time points. The enzymatic digestion of the cell wall to form protoplasts releases cell wall fragments, which can act as damage-associated molecular patterns (DAMPs).

65 TFs were up-regulated and present in the consensus *At-Botrytis* 883-gene network and 61 TFs were down-regulated and present in the same.

A threshold of 0.86 was applied to the edges between these remaining 126 genes of the consensus network. Nodes that only had one incoming edge were also eliminated. This led to the formation of the 70GRN, which consists of 70 nodes and 300 edges. The 70 genes in question are given in Table 4.3.

4.4.3.2 Simulating the Arabidopsis 70GRN with GPDS

In a similar way to learning the DREAM networks in Sections 4.4.1.1 and 4.4.1.3, a GP prior over the functions connecting regulators and targets in the network was used. Hyperparameters were inferred as previously described, this time for the dataset consisting of 70 genes, 24 control time points (where the leaf was uninfected, and water was used instead of *B. cinerea* spores) and 24 infected time points (starting from the addition of a suspension of *B. cinerea* spores onto the leaves as time point 0) and 4 biological replicates. There is no independent dataset to test the accuracy of the model, so a subset of the time series was used as training data. The first 18 out of 24 measurements from 3 out of the 4 biological replicates (control and infected) were used for training the GPDS. The complete fourth biological replicate dataset was used as validation data. It is possible that the covRQard covariance that was best at simulating the DREAM data may not be best at simulating the Arabidopsis gene expression data so the modelling was carried out with multiple kernels. The covariance kernels used in this case are the covSEiso, covSEard, covRQiso, covRQard, covSEiso + covSEiso, covSEiso * covSEiso, covSEiso + LIN, and the covSEiso * LIN kernel.

The only gene that is fixed, as it has no regulators, is gene 67. Predictions on this WT dataset are shown in Figures 4.17-4.21. MSE values for all the kernels are shown in Table 4.4. The MSE for the Arabidopsis genes is larger than for the DREAM

AT1G02170	G1	AT2G33710	G24	AT3G62420	G47
AT1G09530	G2	AT2G37430	G25	AT3G57480	G48
AT1G19180	G3	AT2G38250	G26	AT3G12910	G49
AT1G22190	G4	AT2G38470	G27	AT3G47500	G50
AT1G25440	G5	AT2G40140	G28	AT3G53600	G51
AT1G25550	G6	AT2G40740	G29	AT4G01250	G52
AT1G27720	G7	AT2G43000	G30	AT4G12040	G53
AT1G27730	G8	AT2G44430	G31	AT4G17500	G54
AT1G28370	G9	AT2G44840	G32	AT4G17880	G55
AT1G42990	G10	AT2G24570	G33	AT4G30080	G56
AT1G54140	G11	AT2G47520	G34	AT4G36990	G57
AT1G62300	G12	AT2G32020	G35	AT4G00270	G58
AT1G67100	G13	AT3G05200	G36	AT4G20380	G59
AT1G70920	G14	AT3G13040	G37	AT5G05610	G60
AT1G71030	G15	AT3G19580	G38	AT5G24110	G61
AT1G71520	G16	AT3G23230	G39	AT5G47230	G62
AT1G72310	G17	AT3G23250	G40	AT5G47370	G63
AT1G61960	G18	AT3G28210	G41	AT5G49520	G64
AT1G80840	G19	AT3G46620	G42	AT5G50570	G65
AT2G03340	G20	AT3G49530	G43	AT5G56960	G66
AT2G04880	G21	AT3G52800	G44	AT5G57660	G67
AT2G23320	G22	AT3G55560	G45	AT5G59820	G68
AT2G24650	G23	AT3G61890	G46	AT5G64340	G69
				AT5G66320	G70

Table 4.3: The Arabidopsis Gene Identifiers (AGIs) for the genes in the 70GRN.

dataset genes as the range of gene expression values is larger for the Arabidopsis genes (5 - 17) compared to the DREAM genes (0 - 1). The time series is also longer for the Arabidopsis genes - 48 time points compared to 21. While all data is Z-scored for the purpose of GPR, the MSE is calculated based on the unnormalised data.

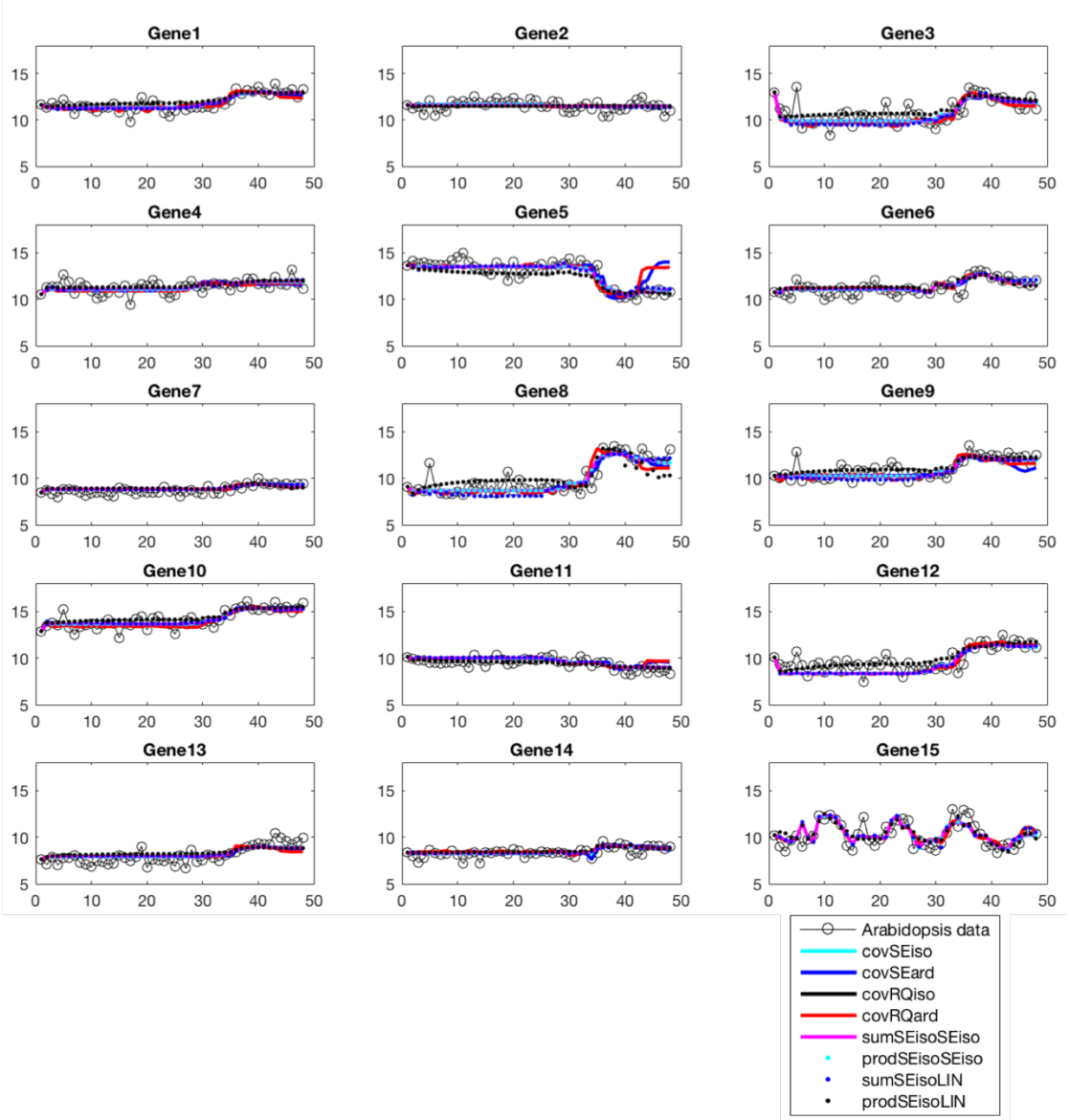


Figure 4.17: GPDS predictions for the WT *At-Botrytis* dataset, genes 1-15. The average experimental gene expression data of 4 biological replicates is plotted in solid black line with circular markers. Predicted gene expression data from GPDS simulations with 8 different covariances is also plotted. The first 24 time points are from uninfected leaves, and time points 25-48 are from infected leaves, and these are treated as one continuous dataset.

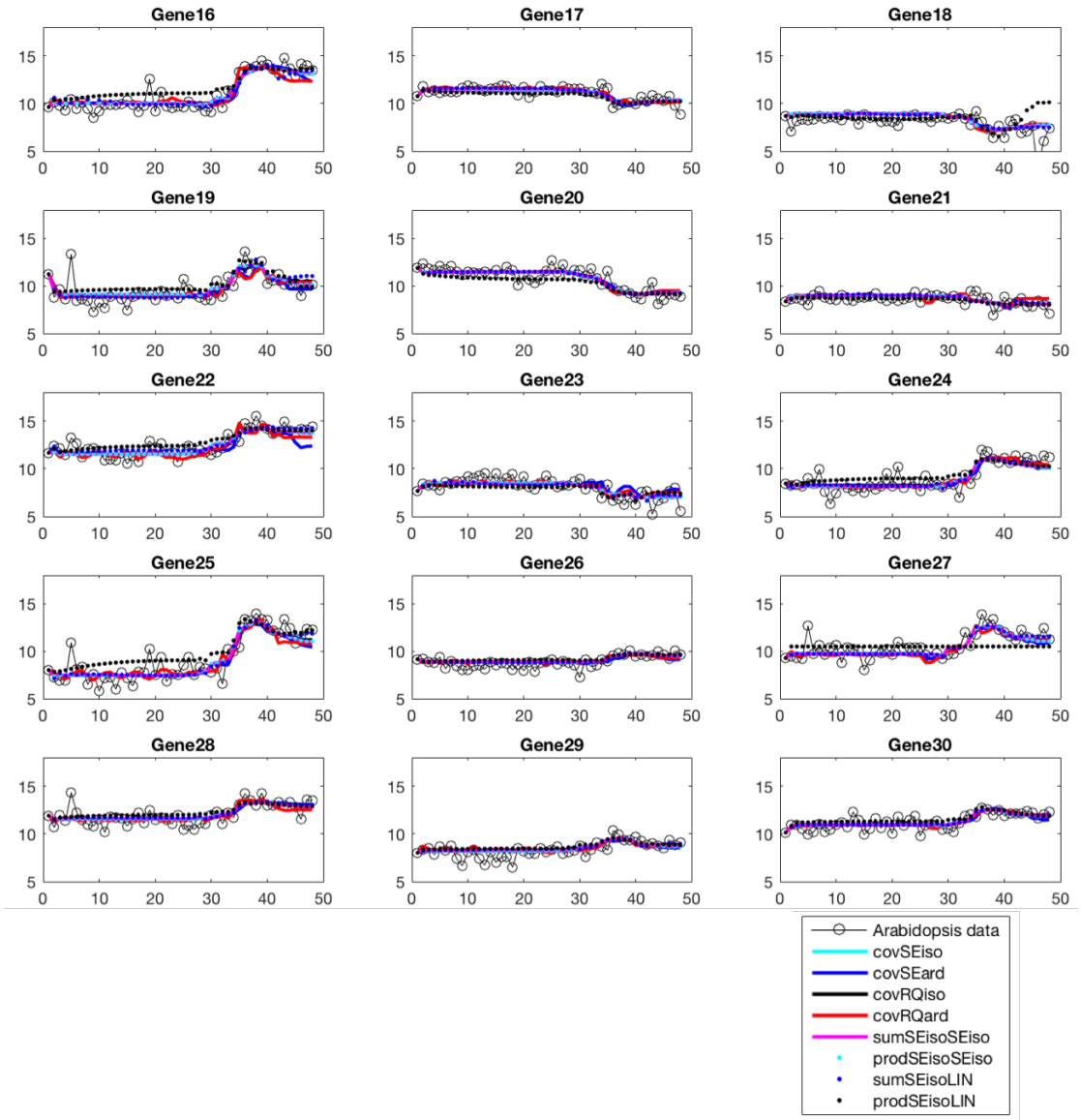


Figure 4.18: GPDS predictions for the WT *At-Botrytis* dataset, genes 16-30. The average experimental gene expression data of 4 biological replicates is plotted in solid black line with circular markers. Predicted gene expression data from GPDS simulations with 8 different covariances is also plotted. The first 24 time points are from uninfected leaves, and time points 25-48 are from infected leaves, and these are treated as one continuous dataset.

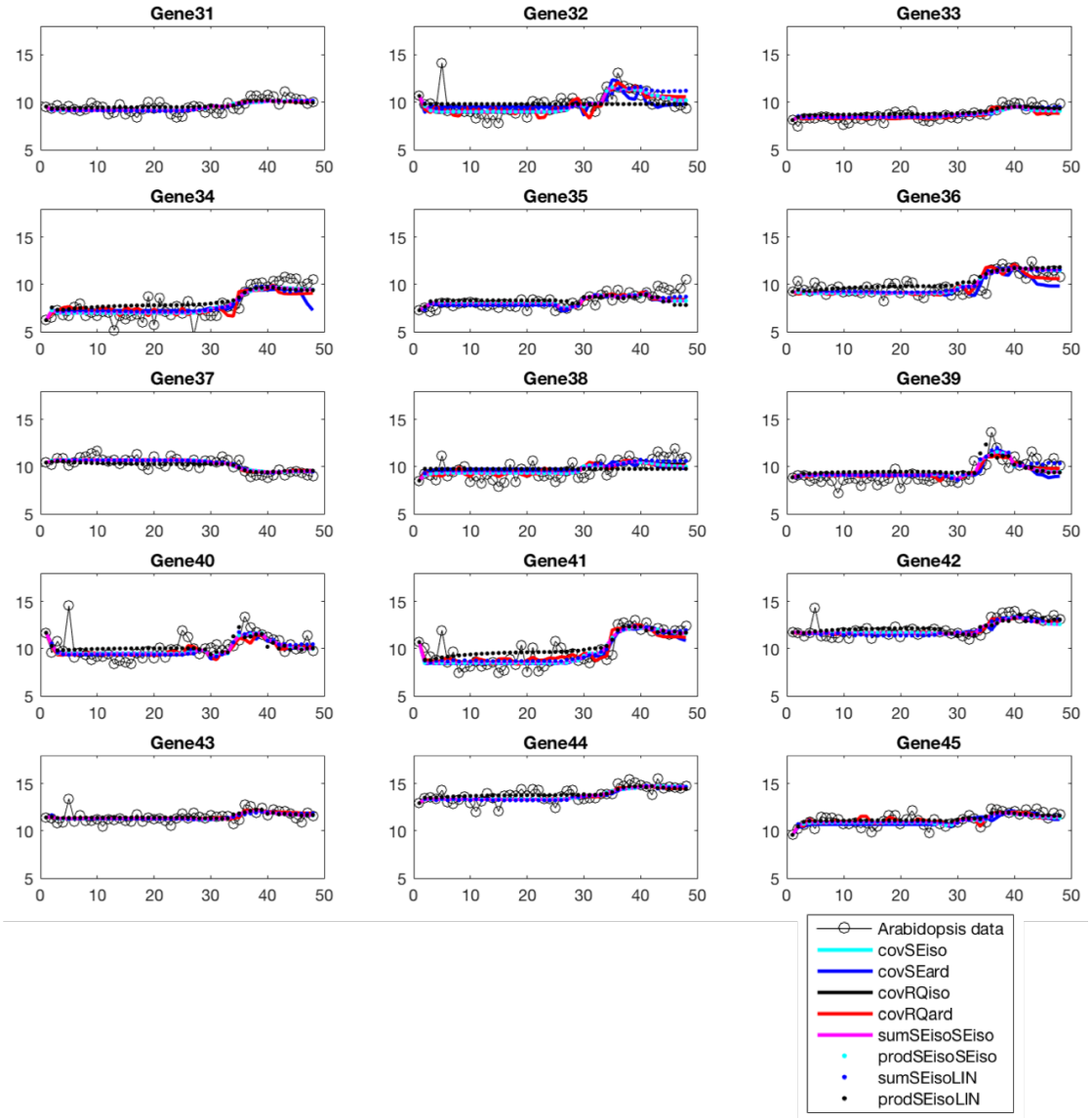


Figure 4.19: GPDS predictions for the WT *At-Botrytis* dataset, genes 31-45. The average experimental gene expression data of 4 biological replicates is plotted in solid black line with circular markers. Predicted gene expression data from GPDS simulations with 8 different covariances is also plotted. The first 24 time points are from uninfected leaves, and time points 25-48 are from infected leaves, and these are treated as one continuous dataset.

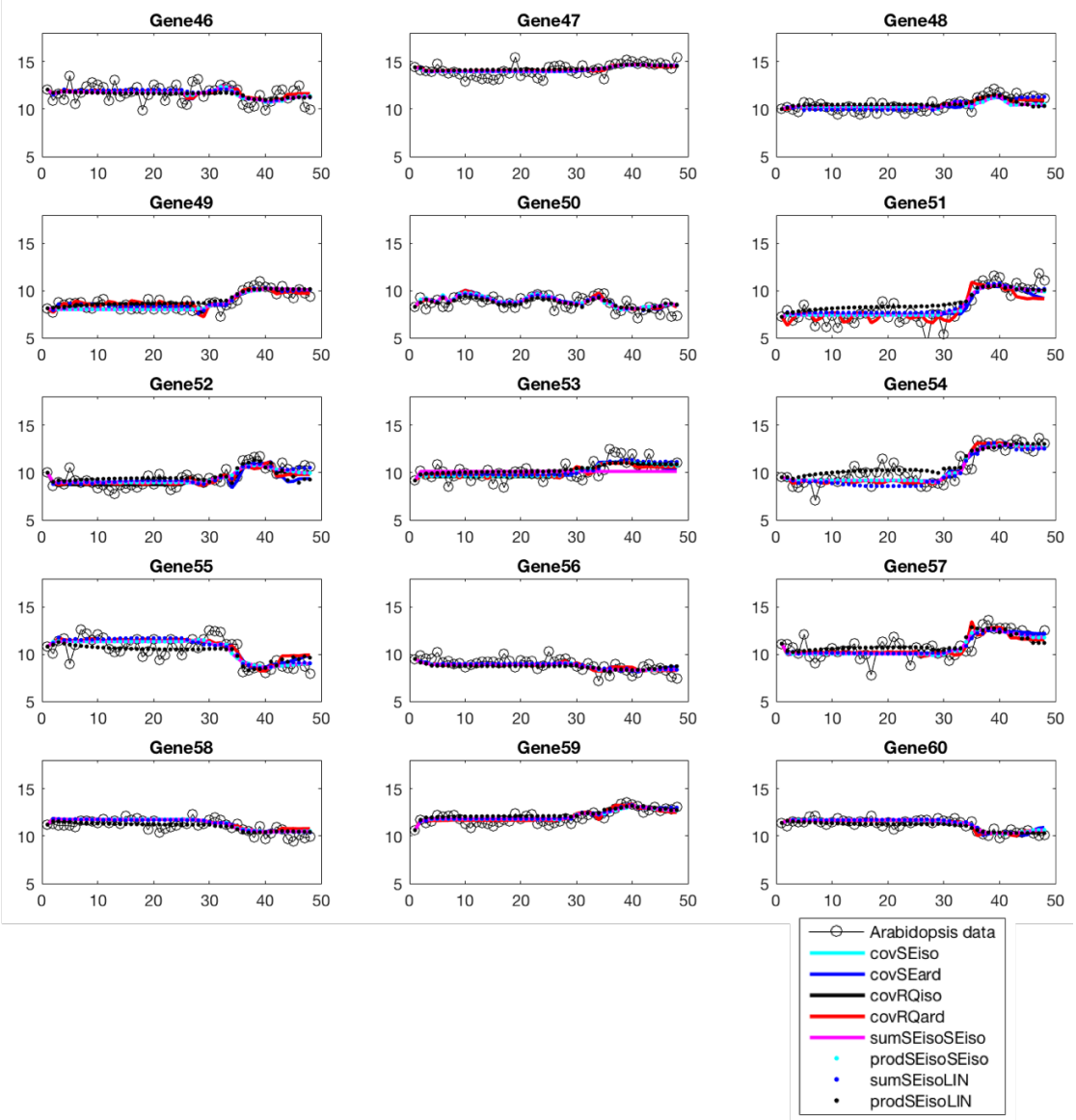


Figure 4.20: GPDS predictions for the WT *At-Botrytis* dataset, genes 46-60. The average experimental gene expression data of 4 biological replicates is plotted in solid black line with circular markers. Predicted gene expression data from GPDS simulations with 8 different covariances is also plotted. The first 24 time points are from uninfected leaves, and time points 25-48 are from infected leaves, and these are treated as one continuous dataset. Gene 67 is fixed.

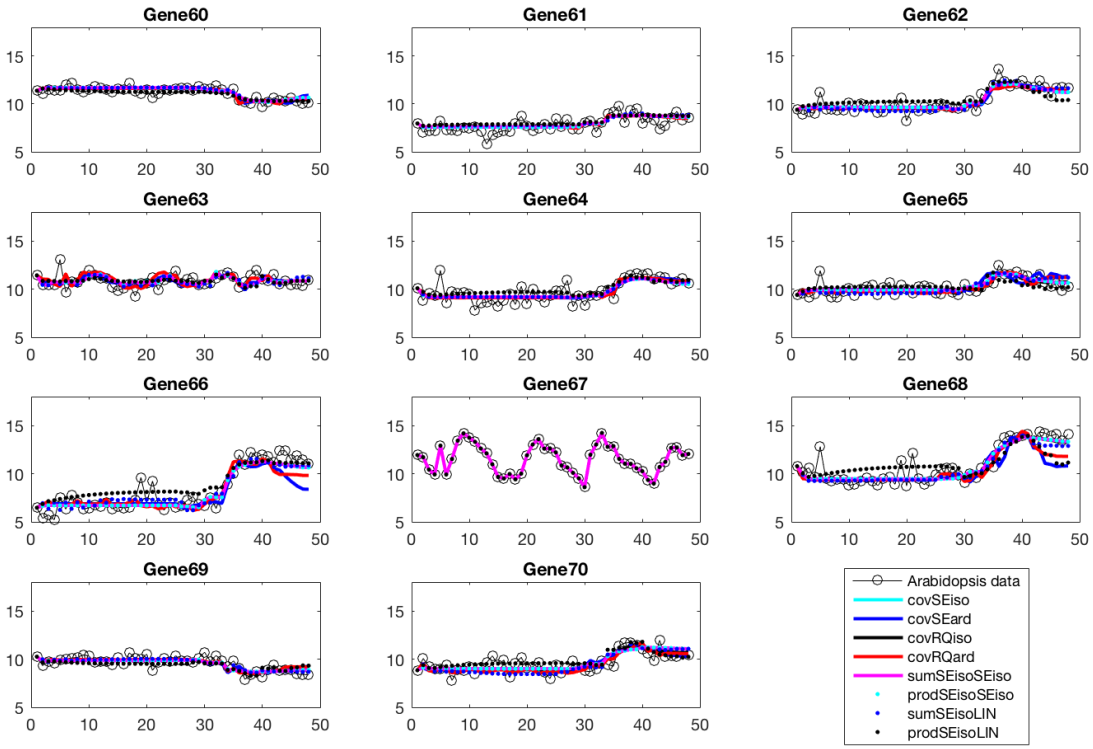


Figure 4.21: GPDS predictions for the WT *At-Botrytis* dataset, genes 61-70. The average experimental gene expression data of 4 biological replicates is plotted in solid black line with circular markers. Predicted gene expression data from GPDS simulations with 8 different covariances is also plotted. The first 24 time points are from uninfected leaves, and time points 25-48 are from infected leaves, and these are treated as one continuous dataset. Gene 67 is fixed.

Table 4.4: MSE for GPDS predictions made with various kernels on the 70GRN validation dataset. The smallest MSE was achieved using the isotropic rational quadratic kernel, followed by the isotropic squared exponential kernel.

Kernel	MSE
covSEiso	8.60
covSEard	11.99
covRQiso	8.42
covRQard	12.15
covSEiso + covSEiso	8.81
covSEiso * covSEiso	8.58
covSEiso + LIN	9.83
covSEiso * LIN	18.79

AGI	Gene no.	Gene Name	Phenotype	Source
AT2G03340	G20	*WRKY3*	Positive	Unpublished data
AT2G38470	G27	*WRKY33*	Positive	Unpublished data
AT4G17500	G54	*At-ERF1*	Negative	Foo *et al.*, 2018
AT5G05610	G60	*AL1*	Positive	Unpublished data
AT5G47230	G62	*ERF5*	Positive	Moffat *et al.*, 2012

Table 4.5: Known phenotypes for the genes in the 70GRN.

4.4.3.3 Re-wiring and optimising the Arabidopsis 70GRN

The covRQiso kernel was chosen to simulate rewirings, as it had the lowest MSE value. Before this was done, the GPDS model was trained on the whole time series data, rather than just 3/4, as done above. The full set of 4761 rewirings (all pairwise rewirings, minus redundant ones and minus rewirings from gene 67 as the source gene, as it does not have any regulators) were then carried out. 5 genes in the 70GRN have associated phenotypes. The objective is to optimise the level of these genes, using the knowledge we have of their effects on *B. cinerea* infection. Positive regulators of defence are are genes that have a positive effect on the defence response in Arabidopsis; knocking these out makes the plant more susceptible. Negative regulators of defence are genes that have a detrimental effect on the defence response; knocking these out makes the plant more resistant to *B. cinerea* infection. The phenotypes of these genes are presented in Table 4.5.

The differential expression of a gene after rewiring is classified as either 'increased' or 'decreased', depending on the average change in the last 12 time points of the simulation:

$$f(x) = \begin{cases} \text{increase} & \text{if } (\frac{1}{19} \sum_{t=36}^{t=48} Z_t^x - Y_t^x) > 0.5 \\ \text{decrease} & \text{if } (\frac{1}{19} \sum_{t=36}^{t=48} Z_t^x - Y_t^x) < 0.5 \end{cases} \tag{4.13}$$

where Z_t^x is the simulated expression of the rewired gene x at time t and Y_t^x is the simulated expression of gene x at time t in the WT network. If the average change in the level of the gene is smaller than 0.5, then it is not counted as differentially expressed.

Given the known phenotypes of genes in the 70GRN during *B. cinerea* infection (see Table 4.5), the criteria for the ideal gene expression are:

- Increase the value of G20
- Increase the value of G27

- Decrease the value of G54
- Increase the value of G60
- Increase the value of G62
- Maintain the values of the remaining genes

Unlike the rewirings simulated for the DREAM network, the gene expression changes following the rewiring of the 70GRN were very limited. Out of all the rewirings tested, most (∼2905) resulted in a single gene being differentially expressed according to the criteria specified in Equation 4.13. This was the gene that was directly targeted by the rewiring. 1723 rewirings resulted in 2 DEG, 4 rewirings resulted in 3 DEG and 58 rewirings resulted in no DEG. Only 1 rewiring resulted in 6 genes being DE (the maximum number of DEG); the DEGs from this rewiring are plotted in Figure 4.23. Out of the 6 DEG in the rewiring of G8 with the promoter of G65, one is G8 itself, and the rest are direct targets of G8. 131 rewirings resulted in the levels of either G54 being decreased of G20 or G60 being increased (an example is plotted in Figure 4.22). All these involved direct rewirings of G20, G54 and G60, or rewiring G8, which decreased the levels of G54.

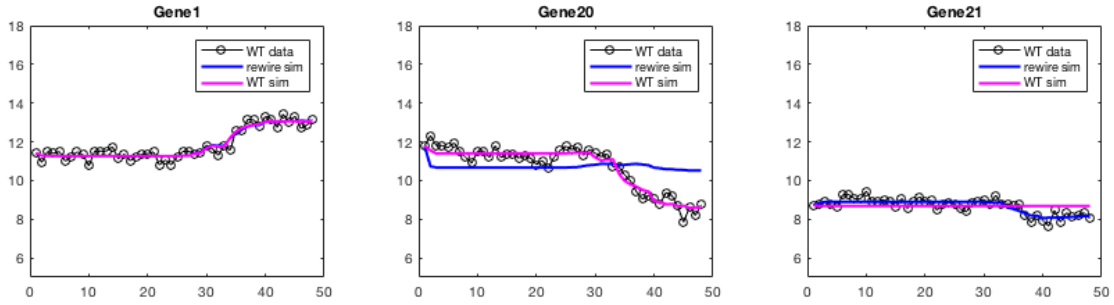


Figure 4.22: GPDS model simulations for G1, G20 and G21 in the rewiring of G20 with the regulatory region of G1. G20 and G21 were DEG in this rewiring, although a closer look at G21 shows that it does not differ significantly from the Arabidopsis data upon rewiring, but the WT simulations are quite inaccurate, leading to the impression that the rewired version of the gene is DE. The GPDS model simulations for the WT network are shown in blue, and the predictions for the rewired network are shown in pink. Solid black line with circular markers: average microarray measurements for the WT network for 4 biological replicates. First 24 time points are from uninfected leaves, time points 25-48 are from infected leaves.

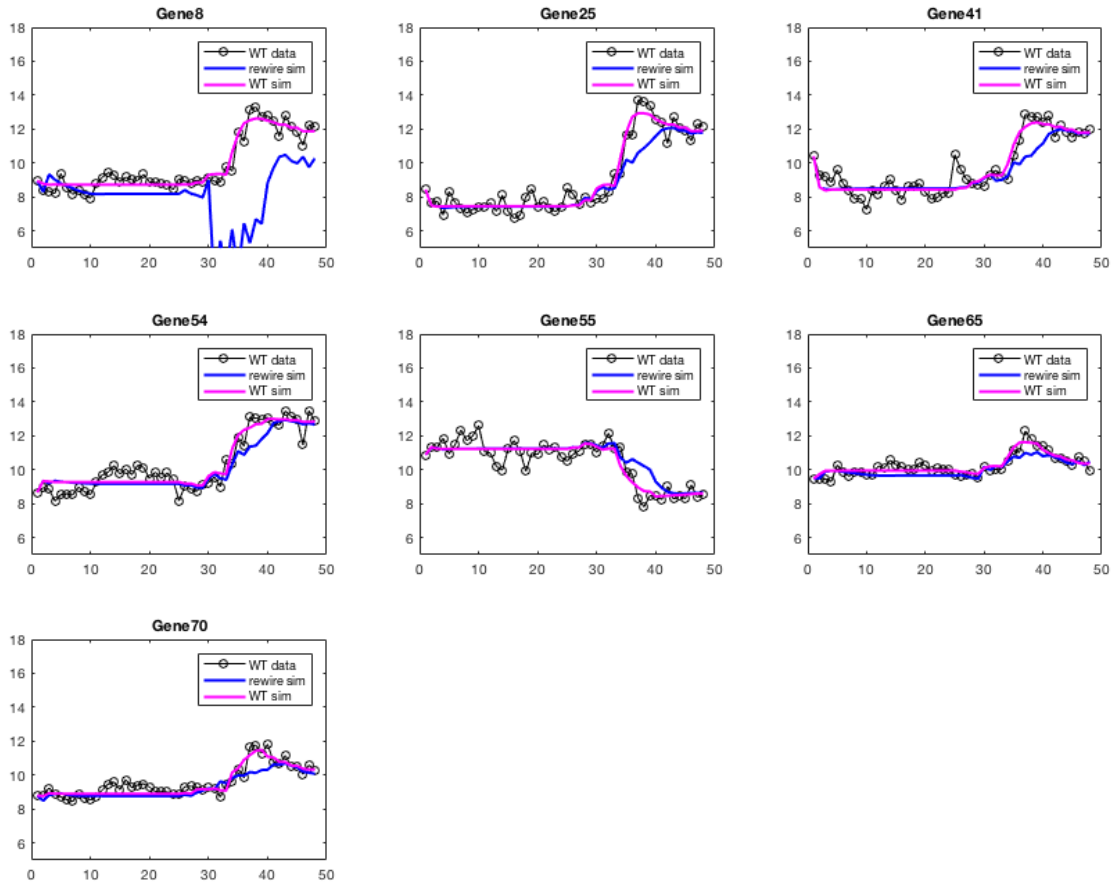


Figure 4.23: GPDS model simulations for G8, G25, G41, G54, G55, G65 and G70 in the rewiring of G8 with the regulatory region of G65. These genes have been plotted as they are all DE following this rewiring (other than G65, which was not classified as DE). The GPDS model simulations for the WT network are shown in blue, and the predictions for the rewired network are shown in pink. Solid black line with circular markers: average microarray measurements for the WT network for 4 biological replicates. First 24 time points are from uninfected leaves, time points 25-48 are from infected leaves.

4.5 Discussion

Gaussian process dynamical systems for modelling and exploring unknown functions is a powerful tool, but have rarely been employed in simulating gene regulatory networks (Penfold and Wild, 2011). This chapter has explored their functionality for such purposes, especially in the case of large networks, and in making predictions for GRNs under new conditions. While the simulations for the validation dataset of the DREAM10 and DREAM100 networks were good, GPDS seem less suited to make rewiring predic-

tions for the Arabidopsis network. For the rewiring of the DREAM10 network, the GPDS made predictions that were the opposite to the real gene expression changes. One reason for this could be due to the different degradation rates of genes. It is possible that the training data provided was not sufficient for the GPDS to capture this detail, and therefore the simulations are less accurate when these rates are assumed to be the same for all genes.

Some genes in the network have no (known) regulators, such as genes 1, 8 and 9 in the DREAM10.1 network (Figure 4.10). These are harder to accurately simulate using GPDS as their profiles are more dependant on the gene itself (such as its degradation rate) rather than the levels of some regulator. This issue was dealt with by "fixing" the levels of these genes - the real gene expression values were always available to the GPDS during the simulation process, unless they were being rewired. This assumes that were this a real network, the expression data for such genes will be available from biological experiments, which is not always the case. While it may not be feasible to fix certain genes with no regulators in the network, as obtaining such biological data is costly and time consuming, it may be enough to just measure the gene expression levels of certain genes at one or a few time points and use that to "pin down" the GP. Further work is needed to determine how many genes and how many time points need to be known in order to achieve greater accuracy. Alternatively the genes without known regulators could be simulated using many different random starting points.

Different covariance kernels were trialled in the GPDS simulations in order to determine whether some were more suited to learning the relationships in GRN. This did turn out to be the case - change-point kernels and kernels that combined two covariance functions were not as accurate as the simpler squared exponential and rational quadratic kernels. Interestingly, the kernels with the smallest MSE (for the DREAM network simulations) were always the ones with individual length scale hyperparameters (Automatic Relevance Determination) for each of the regulators. This implies that the way TFs affect genes can vary with the TF in question and should be taken into account in modelling GRNs. This is well known in plant TFs; in addition, TFs can also act as positive regulators for some genes and negative regulators for others (Dombrecht *et al.*, 2007). The idea of applying different covariance kernels to a regression problem in order to construct the most accurate model of a dataset was explored by Lloyd *et al.* (2014). However, their 'automated statistician' was constructed to extrapolate existing datasets, rather than making predictions for the system under new conditions.

Unlike the DREAM training data, which contains time series for a network exposed to different perturbations, there is a lack of high resolution data for biological networks under different conditions. Realistically, the GPDS model should be able to perform well given limited training data if one is to make predictions that can be tested out *in planta*. While GPDS benchmarking of the DREAM networks was carried out using all available resources (such as all training data, certain fixed genes), it would be helpful to determine how the GPDS rewiring performance varies when training data is withheld. For instance, when simulating the rewirings with GPs for the DREAM networks, it is assumed that the starting point is known for all genes, which is an un-

realistic scenario. Rather, simulations should be started at multiple random points to reflect this uncertainty. It will also be useful to expand the modelling so that it is able to incorporate other kinds of data, such as single point knock-out data or protein data, as this is more readily available than high resolution time series at different perturbations. This has been done by Äijö and Lähdesmäki (2009), who incorporated data from KO experiments and steady-state measurements into their ODE + GP model of gene expression. Essentially, an extra data point was added for the knocked-out gene y, measured at steady state: $y = 0$ at $dy/dt = 0$ (their model inferred rates similar to the model in Equation 4.10). This approach would work well in combination with the single time point microarray data that is already available (see Section 2.5.2). Finally, using a warped GP (Snelson *et al.*, 2004) will be helpful in preventing gene levels from decreasing below 0 or increasing beyond a reasonable level.

The rewiring predictions made for the Arabidopsis TF network were very similar to the WT simulations. In most cases, the only DEGs in the simulations were the gene that was the subject of rewiring - no gene downstream of it was significantly affected. This is perhaps due to the fact that the GP model was trained on limited data with two conditions - infected and not infected, compared to the DREAM models, which were trained on time series from 20 different conditions (perturbations). Therefore the GPDS model was only able to observe very limited dynamics of the Arabidopsis genes, which makes it harder to extrapolate different scenarios, like switching regulators. In other words, the existing observations are not sufficient to map the GP onto the phase space, as demonstrated in Figure 4.4. It is also probable that in selecting only the genes that behave in a similar way in protoplasts and Arabidopsis plants, a number of important regulators and connections were left out of the network. The network threshold that was chosen for this purpose may also have been too stringent, so influential connections may have been left out. Due to this missing data, the GPDS models may not be able to learn the functions describing the expression of genes. Additionally, the starting network, the *At-Botrytis* model, may contain many small inaccuracies which can be overlooked when using ODEs to model small subnetworks such as the 9GRN, but which add up and are detrimental when modelling the larger 70GRN. Ultimately, by modelling only TFs that are similarly DE in protoplasts and leaves leads to too many assumptions and genes left out of the modelling, defeating the purpose of a systems-level model. In the future, this problem can be addressed by acknowledging these differences and separately modelling the networks in protoplasts and leaves.

Biological networks are famously robust to interference within their network structure (Isalan *et al.*, 2008). Hence a single rewiring may not achieve the desired response due to dampening of the changes. Multiple rewirings may be necessary, and can be modelled with our system and tested out using methods such as Golden Gate (Engler *et al.*, 2008) to create multiple promoter-gene constructs in a single transformation. Golden Gate easily allows the assembly of DNA modules onto a single plasmid. If multiple rewirings are simulated, it may not be practical to do so for every possible combination. Instead, methods like approximate Bayesian computation (Liepe *et al.*, 2014) can be used to scan for the best network structure that exhibits certain gene expression

patterns. Approximate Bayesian computation can operate a model selection framework that can identify the best model that fits a certain dataset (such as the optimised or desired gene expression data for a system). The best network can be narrowed down in a step-like fashion with a sequential Monte Carlo framework, where each at each step, models are weighed by their ability to explain the data. The Approximate Bayesian computation algorithm introduced by Toni *et al.* (2009) uses ODEs for modelling; however, this could be replaced with GPDS to allow the simulation of larger systems.

In order to ensure the GPDS performance is robust to different perturbations and rewirings, simulations were carried out for thousands of rewirings, and five different network perturbations. This has naturally resulted in a lot of gene expression data, most of which was examined by hand. While the MSE is a useful tool for roughly estimating how well the GPDS predicts data, it is a summary of all the genes in the network and can hide systematic weaknesses - such as certain expression profiles that are poorly simulated, if the simulations on the rest of the genes are accurate. This particular project would benefit from better data visualisation and summary statistics describing how well the GP model performed on certain problems. For instance: did it fail to replicate a dynamic profile? Did it revert to simulating the gene based on the WT network rather than the rewired network? Did it fail to take into account perturbations or degradation rates during rewiring? Identifying the systematic errors and when they occur will help in developing strategies to tackle them.

4.6 Materials and Methods

The DREAM datasets are described in the Materials and Methods section of Chapter 2.

4.6.1 GeneNetWeaver

GeneNetWeaver (GNW), a simulation tool for *in silico* networks, was downloaded from `http://gnw.sourceforge.net`. GNW extracts network information specified in SBML in .xml files. Details on how to download the SBML files detailing the ODE models and parameters for the DREAM4 models are given in Section 2.5.1. In order to simulate rewiring of Gene X with Gene Y, the SBML files were modified by replacing the details of the synthesis reaction of Gene X with the synthesis reaction of Gene Y. This included replacing the regulators and the associated parameters denoting the strength of the regulation. The rewired files were then uploaded to GNW and timeseries were generated using the model of noise found in microarrays, and the same perturbations as those applied to the wild-type, un-rewired network.

4.6.2 ANOVA

1-way ANOVAs were performed for the mean-squared error (MSE) values from simulation produced by using different Gaussian process kernels. 30 simulations were

generated using each kernel, the MSE value calculated using the following formula and the ANOVA performed using the MATLAB function *anova1*.

4.6.3 Gaussian process dynamical systems modelling

Code for modelling using GPDS is given in Dataset G.

4.6.3.1 Inferring hyperparameters

In order to infer hyperparameters for a Gaussian process, a minimisation function using conjugate gradients was used to find the best hyperparamemeter values that described the gene expression, given certain regulators. The training data consisted of training inputs (the gene expression of the regulators at time points 1 to $t-1$) and training outputs (the gene expression of the target gene at time points 2 to t). The training data was normalised so that the expression levels of each gene had zero mean and unit variance. The minimisation function was run over 3000 steps with initial guesses for all hyperparameters of 1.

In addition to the expression of regulators of a gene being used as training inputs, the perturbation is also assumed to be a regulator. For the DREAM networks, perturbations are applied to certain genes for the first half of the time series and removed for the second half of the time series. The magnitude of the perturbations is given in Dataset F. For the Arabidopsis networks, the growth of *B. cinerea* is used as the perturbation parameter and is applied to all genes. The relative tubulin expression from Windram *et al.* (2012) Figure 2B is used as a proxy for *B. cinerea* growth. The perturbation experienced by genes is 0 for the control treatment.

4.6.3.2 Covariance functions

The Gaussian processes functions were implemented using the GPML toolbox from `http://www.gaussianprocess.org/gpml/code/matlab/` based on the algorithms described in Rasmussen and Williams (2006), and the change-point functions were obtained from Lloyd *et al.* (2014). A mathematical description of the covariance functions is given in Appendix B. The number of hyperparameters depended on the covariance function used. The kernels used to learn and simulate gene expression, and their associated hyperparameters, are:

- covSEiso - isotropic squared exponential covariance function - length scale, signal variance, noise variance
- covSEard - squared exponential covariance function with Automatic Relevance Determination (ARD) - length scale for each regulator, signal variance, noise variance
- covRQiso - isotropic rational quadratic covariance function - length scale, scale mixture parameter alpha, signal variance, noise variance

- covRQard - rational quadratic covariance function with ARD - length scale for each regulator, scale mixture parameter alpha, signal variance, noise variance
- covSEiso + covSEiso - composite covariance function made from the sum of two isotropic squared exponential covariance functions - length scale, signal variance for each of the two kernels; noise variance
- covSEiso * covSEiso - composite covariance function made from the product of two isotropic squared exponential covariance functions - length scale, signal variance for each of the two kernels; noise variance
- covSEiso + LIN - composite covariance function made from the sum of an isotropic squared exponential covariance function and a linear function - length scale, signal variance for covSEiso kernel, the linear kernel has no hyperparameters; noise variance
- covSEiso * LIN - composite covariance function made from the product of an isotropic squared exponential covariance function and a linear function - length scale, signal variance for covSEiso kernel, the linear kernel has no hyperparameters; noise variance
- CP Const + Const - change-point kernel using two constant covariance functions - location of the change point, steepness, (scalar parameter 1, scalar parameter 2) * number of regulators; noise variance
- CP Const + covSEiso - change-point kernel using a constant and an isotropic squared exponential covariance function - location of the change point, steepness, (scalar parameter 1, length scale, signal variance) * number of regulators; noise variance

4.6.3.3 Simulating gene expression

In order to simulate gene expression, the same training data is used as for inferring hyperparameters. The initial true values of each gene's regulators are also given. The *gp* function from GPML is then used to infer the value of the gene of interest at time point 2 given the training data, the values of its regulators at the previous time point, the hyperparameters of the covariance kernel and exact inference for a GP with Gaussian likelihood. This is repeated for every gene in the network, with the exception of genes whose levels are fixed ie their true values are known throughout the simulations. The genes that do not have any regulators in the network are fixed. This whole process is then repeated for time point 3, and all subsequent time points, with the updated vales of the regulators inferred from the previous time point. The MSE of a simulation with regards to the validation dataset is calculated with Equation 4.11 once the simulations have finished. Boxplots for the MSE were created using the MATLAB function *boxplot*.

4.6.3.4 Model Re-wiring

Simulating gene expression from a rewired network is very similar to simulating gene expression from the WT network. The difference is that the regulators for the target gene of the rewiring are exchanged with regulators of the source gene. The training data for the target gene are also replaced with the training data of the source gene. The same is done for the hyperparameters of the target gene, except for the hyperparameters associated with the perturbation, which remain unchanged.

4.6.3.5 Model Optimisation

Each gene in each rewiring simulation is given a classification for its response to perturbation: 'increase', 'decrease' or 'unchanged'. 'Increase' means the average gene expression levels after perturbation are increased by more than 0.1 compared to pre-perturbation levels. 'Decrease' means the average gene expression levels after perturbation are decreased by more than 0.1 compared to pre-perturbation levels. 'Unchanged' means the average gene expression levels after perturbation are within 0.1 of the pre-perturbation levels.

Chapter 5

A high-throughput system for testing re-wiring in Arabidopsis protoplasts

5.1 Aims

While GPR modelling from the previous chapter was not refined enough to make good predictions for rewiring of Arabidopsis genes, rewiring can still be carried out *in vivo* based on hypotheses formed from biological experiments to enhance the Arabidopsis defence response to *B. cinerea*. In this chapter, rewirings are chosen so that enhancers of the defence response which are downregulated during infection can be upregulated by attaching promoters from strongly up-regulated genes to their coding regions. A protoplast system is evaluated for its ability to capture the effects of rewiring in plants.

- Establish the feasability of protoplasts as a system for testing out strategies for improving the Arabidopsis defence response to *B. cinerea*.
- Determine the amount of overlap in the leaf protoplast response to chitin and the leaf response to *B. cinerea* infection.
- Construct plasmids of promoter-gene fusions for rewiring the transcription factor (TF) networks pertinent to *B. cinerea* stress in Arabidopsis.
- Compare and contrast the effects of rewiring protoplasts and Arabidopsis plants (stable transformants).

5.2 Acknowledgements

Library preparation for RNA sequencing was performed by Dr. Sally James from the Technology Facility at the University of York Department of Biology. The Oxford Genomics Centre at the Wellcome Centre for Human Genetics was responsible for the

generation and initial processing of RNA sequencing data. All other work was carried out by myself.

5.3 Introduction

5.3.1 Protoplasts as a versatile high-throughput system for molecular genetics

Protoplasts allow the study of gene function, gene regulation, the appearance of cellular structures and the application of gene editing in a high throughput manner (Marx, 2016). Protoplasts are isolated from plant tissue by removing the cell wall through mechanical or enzymatic means (Wu *et al.*, 2009a; Yoo *et al.*, 2007). While protoplasts cannot reproduce and are therefore generally used for transient assays, they can form calluses and regenerate whole plants (Bourgin *et al.*, 1979; Shepard and Totten, 1977; Yamada *et al.*, 1986). This property in particular is exploited for CRISPR-Cas9-induced multi-allelic mutagenesis without the integration of the DNA in the plant genome (Andersson *et al.*, 2017; Shan *et al.*, 2013).

Leaf protoplasting leads to protoplasts from a mixture of tissues; however, tissue specific cells can be isolated using tissue-specific fluorescent markers and fluorescent assisted cell sorting (Sparks and Benfey, 2017). Protoplasts have been a useful tool for deciphering the plant stress response, such as studying the function of Mitogen-activated protein kinase kinases and their network interactions in cold and salt stress (Teige *et al.*, 2004), to understand the response of abscisic acid (ABA)-responsive element binding proteins to ABA (Uno *et al.*, 2000), and to visualise the localisation of putative zinc-finger transcription factors and assess their transcriptional activity under abiotic stress (Sakamoto *et al.*, 2004).

Advances such as the development of a 96 well plate-based transcription factor (TF) transactivation system (Wehner *et al.*, 2011) and the use of liquid handling robots make protoplasts a very attractive screening system. This allowed the identification of regulatory links to promoter::luciferase constructs. In synthetic biology, protoplasts have been used in the work of Schaumberg *et al.* (2015) to characterise promoter strength and produce a library of > 100 tunable synthetic repressors.

However, one key drawback to this system is the stress and transcriptional changes caused by stripping away the cell wall of Arabidopsis cells (Gifford *et al.* (2008), Birnbaum *et al.* (2005)). This can mask the effects of treatments as genes may already be differentially expressed (DE) due to the protoplasting itself. A second drawback is that whole-plant phenotypes, such as lesion size or other responses to infection, cannot be observed in protoplast cultures.

5.3.2 The use of microbial-associated molecular patterns for eliciting the defence response

Protoplasts cannot actually be infected with pathogens, so microbial-associated molecular patterns (MAMPs) or damage-associated molecular patterns (DAMPs) can be

used instead to induce an immune response (Ranf *et al.*, 2011). For instance, chitin can be used in lieu of *Botrytis cinerea* and flg22 can be used in lieu of *Pseudomonas syringae* infection. The recognition of flg22 by the plant immune system and the subsequent responses have been widely characterised (Felix *et al.* (1999), Gómez-Gómez and Boller (2000), Zipfel *et al.* (2004), Benschop *et al.* (2007), Denoux *et al.* (2008)). Flg22 is a conserved N-terminal peptide of 22 amino acids in flagellin, the building block of the bacterial flagellum. This region is conserved in a large range of Gram-positive and Gram-negative bacteria (see Figure 2 in Felix *et al.* (1999)). The flg22 epitope is recognised by the Arabidopsis FLAGELLIN-SENSING 2 (FLS2), a transmembrane leucine-rich repeat receptor kinase (LRR-RK) and is sufficient for instigating an immune response (Felix *et al.*, 1999). Binding of FLS2 to flg22 triggers multiple defence signalling pathways, and in particular jasmonic acid (JA) associated processes (Denoux *et al.*, 2008). However, this has also been shown to activate independently of salicylic acid (SA), ethylene (ET) or JA signalling (Ferrari *et al.* (2007), Zipfel *et al.* (2004)). The early response to flg22 treatment includes activation of SA, JA and ET pathways, the upregulation of antimicrobial biosynthetic genes, senescence and SA-mediated secretion pathways (Denoux *et al.*, 2008).

Chitin is a polysaccharide composed of N-acetylglucosamine and chitosan is its deacetylated derivative. Chitin is mainly found in the exoskeleton and internal structures of invertebrates, and is generally extracted from crustacean cells. It is also found in the cell walls of yeast and fungi (Rinaudo, 2006). The chitin MAMP is recognised by a LysM-type receptor-like kinase CERK1 (Miya *et al.*, 2007; Willmann *et al.*, 2011). Both chitin and chitosan can be used to elicit the defence response and prevent post-harvest diseases (Zhang *et al.*, 2011a). Low molecular weight chitosan inhibited citrus fruit decay caused by a number of fungi including *B. cinerea* (Chien *et al.*, 2007), and application of *R. glutinis* with chitin inhibited post-harvest grey mould of strawberries (Ge *et al.*, 2010). Similarly, adding chitin to the culture media of *C. laurentii* enhanced its antagonistic activity against *Penicillium expansum*-derived blue rot in pear (Yu *et al.*, 2008). This priming action of chitin and its derivatives is a result of its ability to induce the defence response in a similar manner to fungi.

5.3.3 Synthetically rewiring networks *in vivo*

Rewiring - replacing the regulatory regions of genes with different regulatory regions - has been shown to provide a viable approach to increasing tolerance to environmental conditions (Isalan *et al.*, 2008) or increasing protein production (Windram *et al.*, 2017).

Transcription factor networks are extremely robust to changes, as shown by the 568 rewirings performed in *E. coli* by Isalan *et al.* (2008). In this study, master TFs, together with σ factors and some downstream TFs had their promoter regions substituted with all the other promoter regions in turn. These constructs composed of a promoter region + transcription factor coding region + *GFP*-coding region were then integrated into the genome, leading to the introduction of new loops, cascades or feedback motifs. Surprisingly, *E. coli* tolerated the rewiring of hub genes which regulate up to 1000 genes

(such as the sigma factor sigma70) as well as non-hub genes and 95% of the transformants were viable. In addition to this, they showed that certain rewirings provided specific fitness advantages in conditions such as heat shock and extended periods at 37°C. Therefore the network structures that are currently present in organisms are not necessarily optimised for all functions and there is cause for exploring alternative configurations.

Further work to characterise profiles of rewired networks in *E. coli* was carried out by Baumstark *et al.* (2015). 85 rewirings led to differential expression spanning 4 orders of magnitude from 0 to 1,000. Rewiring TFs with a larger numbers of targets or strong promoters was more likely to result in a larger numbers of DEGs. They also used rewiring as a means to confirm predicted interactions or discover new regulatory interactions. Surprisingly, as was also observed by Isalan *et al.* (2008), gene expression appears to be regulated by the open reading frame in question as much as by the promoter.

Windram *et al.* (2017) used rewiring to program a certain phenotype in *Pichia pastoris* yeast cells, namely, enhanced production of heterologous proteins. Cassettes consisting of one of 2880 rewiring combinations and a *GFP* reporter were transformed into *P. pastoris*. Clones with enhanced fluorescence were identified and further analysed; yeast with the *CTA1::URE2-l* rewiring event increased the levels of *GFP* and other heterologous proteins. Certain design principles were inferred from the top performing rewirings; namely, the nodes used had higher outdegree, betweeness centrality and lower clustering coefficient and the promoters had higher positions in the regulatory hierarchy and low clustering coefficients.

Rewiring of endogenous genes has been shown to successfully modify phenotypes in *E. coli* and *P. pastoris*, but it has not been applied by the synthetic biology community to Arabidopsis. it is employed here in an endeavour to enhance the defence response in Arabidopsis. Rewiring is carried out in protoplasts, chosen for its high-throughput, cost-effective and swift implementation, and the defence response is elicited using chitin. Rewirings are also performed in Arabidopsis plants and these experiments are compared to establish whether predictions for good rewirings made using protoplasts can be extrapolated to whole plants. Rewired and non-rewired control plant leaves are infected with *B. cinerea* and the size of the lesion is measured to determine whether the rewiring has improved the defence response.

5.4 Results

5.4.1 Determining optimal chitin induction in protoplasts

In order to establish protoplasts as a high-throughput screening system for rewiring, the response of protoplasts to chitin had to be characterised. This is quantified using RNA sequencing (RNAseq) and compared to the response of genes from Arabidopsis leaves infected with *B. cinerea*. Preliminary experiments using qPCR to measure gene expression were first carried out to optimise the experiment conditions.

The chitin concentration and duration of induction were varied to find the conditions that resulted in robust, rapid and observable differential gene expression in proto-

plasts. While protoplasts can remain alive for days in the osmotically stabilised solutions, if they are left for too long, they start secreting pectin to rebuild the cell wall (Marx, 2016).

Protoplasts were extracted as per Yoo *et al.* (2007) from 3- or 4-week old Arabidopsis leaves. The viability of protoplasts was determined with Fluorescein Diacetate stain, which only fluoresces in the presence of an intact cell membrane (Figure 5.1). The number of viable protoplasts immediately upon extraction and within 24 hours of extraction is not significantly different. The protocol liberates $\sim$1.3x10^6 protoplasts from 24 Arabidopsis plants, as measured using a Fuchs-Rosenthal counting chamber. The protoplasts were left to rest overnight (or were transformed, and then left to express the construct overnight) before being induced with chitin.

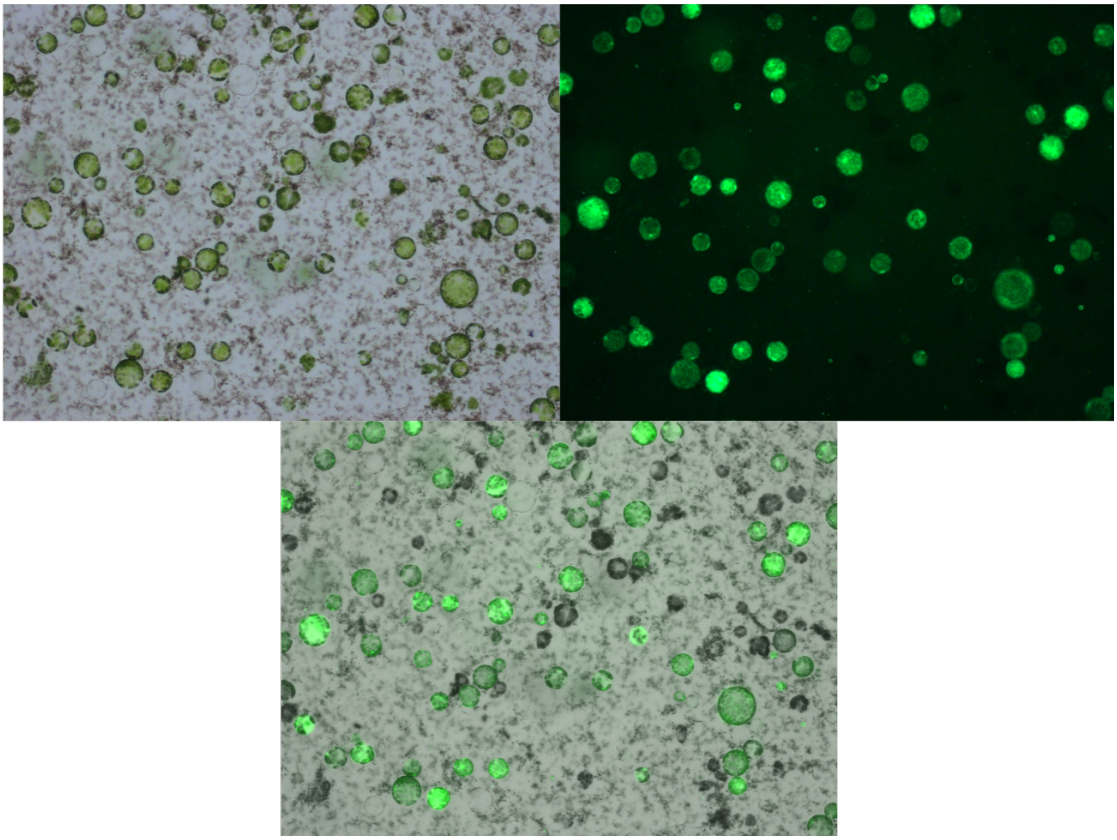


Figure 5.1: Protoplasts stained with Fluorescein Diacetate. Protoplasts imaged in brightfield (top left) and fluorescent channel (top right). Overlay of brightfield and fluorescent channel image (bottom).

The effects of chitin or chitosan has been studied on Arabidopsis seedling gene expression via micrarray technologies (Ramonell *et al.* (2005), Povero *et al.* (2011), Wan *et al.* (2008), Libault *et al.* (2007)). In each case, several hundred genes were DE. A number of those genes were also found DE in *B. cinerea*-infected leaves (Figure 5.2) which

is to be expected as the chitin-triggered response is present in both scenarios. However, many more genes are uniquely upregulated or downregulated during the infection, as expected of a more complex and dynamic process.

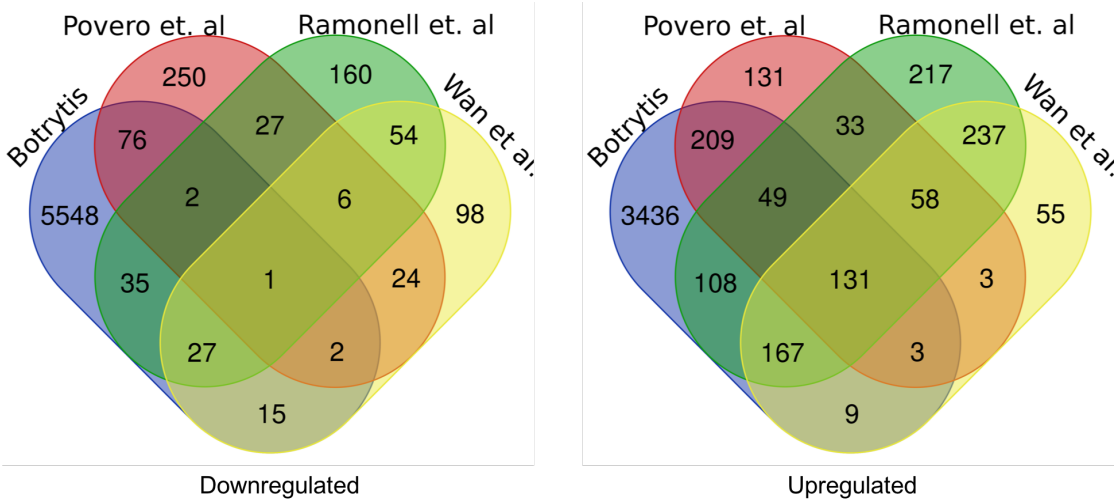


Figure 5.2: Overlap of DEGs from studies where Arabidopsis seedlings have been treated with chitin or its derivatives and *B. cinerea*-infected Arabidopsis leaves (Windram *et al.*, 2012).

Treatment of chitin in Arabidopsis seedlings usually lasted between 15 and 120 minutes in the published works, so these lengths of induction were trialled. Concentrations of 5, 10 and 15 mg/ml were used, corresponding to published concentration (see 5.1). Chitin from shrimp shells was prepared using the method in Millet *et al.* (2010) and diluted in protoplast buffer W1 to form solutions of 5, 10 and 50 mg/l concentrations in the final volume when added to the protoplast suspension. For the control, extra buffer was added to make up the final volume. 1, 2.5 and 5 mg/l of chitosan were also trialled, but these were not as effective at inducing more than a two-fold increase in genes that were expected to be highly upregulated. Additionally, concentrations of 5 ml/l or higher of chitosan had an adverse effect on the protoplasts by encouraging cells to aggregate. From the studies mentioned in 5.1, I looked for highly-expressed genes that could be used as markers for chitin treatment. Although these were carried out in seedlings, it is likely that genes consistently highly upregulated genes in these studies will also behave the same in protoplast treatment with chitin. The two genes selected as markers are AT1G72520 (LOX4) and AT1G56060 (ATHCYSTM3) due to their high differential expression as shown in Table 5.2.

qPCR was performed over two days on these marker genes to find the optimal conditions for differential gene expression. For each condition, 3 biological (a different batch of protoplasts treated with chitin) and 3 technical replicates (3 separate wells in the qPCR plate) were used. The resulting gene expression data were noisy: the variation between biological replicates measured on the same day in the same qPCR plate was

Table 5.1: Summary of all literature studies up to present date carried out on Arabidopsis seedlings measuring gene expression changes to treatment with chitin or chitin derivatives.

Inducer	Time (min)	DEG	Method	Max gene induction (min)	Reference
Chitooctaose 1µM	15, 30, 60, 90, 120	148 out of ~2000	qPCR, Affymetrix array	30, 60	Libault et al., 2007
Chitin 100µg/mL, chito-octamer 1µM	30	5012 out of 23000	Affymetrix array	-	Ramonell et al., 2005
Chitooctaose 1µM	30	890 out of 22000	Affymetrix array	-	Wan et al., 2008
Chitosan 150µg/ml	180	1005 out of 24000	Affymetrix array	-	Povero et al., 2011

Table 5.2: Published fold change in selected marker genes treated with chitin or chitin derivatives.

	FC Povero *et al.*	FC Wan *et al.*	FC Ramonell *et al.*	Average FC
AT1G72520	31.52	41.48	47.01	40.00
AT1G56060	49.13	22.37	12.57	28.02

large (Figure 5.3). 15 minutes was insufficient to generate a response. Chitin treatment (5 and 10 mg/l) for an hour decreased the upregulation seen in genes AT1G72520 and AT1G56060 compared to the equivalent treatments at 45 minutes. Protoplasts treated with 50mg/l of chitin over an hour showed the highest average upregulation - the average expression of AT1G72520 was increased 4.1, 21 and 29-fold relative to the control over the 3 biological replicates; the average expression of AT1G56060 was increased 1.0, 8.2 and 14.3-fold over the control in the 3 biological replicates. The marker genes in protoplasts treated with 10 mg/l of chitin for 45 minutes were induced 3-fold to 10-fold in all 6 biological replicates (over the 2 days). While incubation with 50 mg/l chitin for an hour resulted in the highest fold-change observed and no differential expression in one case, a more conservative chitin concentration of 10 mg/l and incubation for 45 minutes were chosen as the conditions for the follow-on RNAseq experiment. The upregulation was more robust over the latter conditions, and there were more data to support this conclusion.

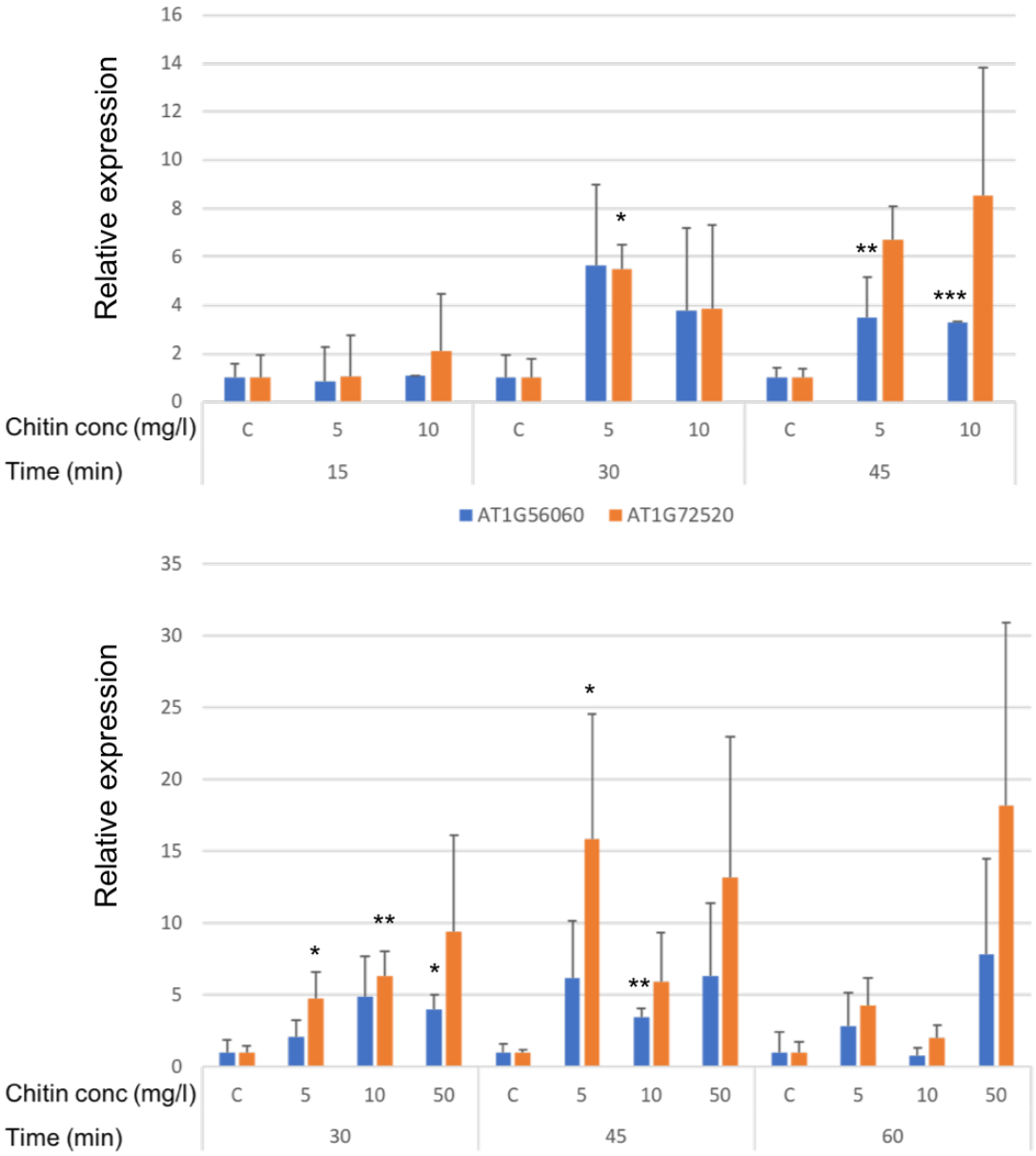


Figure 5.3: qPCR results from two individual experiments of protoplasts treated with chitin. In the top plot, protoplasts were treated with 0 (C - control), 5 and 10 mg/l of chitin for 15, 30 or 45 minutes. In the bottom plot, a different batch of protoplasts were treated with 0 (C - control), 5, 10 and 50 mg/l of chitin for 30, 45 and 60 minutes. RNA was extracted from these protoplasts and the levels of AT1G56060 and AT1G72520 measured with qPCR. The levels of these genes in the different treatments are normalised relative to the control in that particular time point.

Figure 5.3: *(previous page)*: * represents p≤0.05, ** represents p≤0.01, *** represents p≤0.001 in a two-tailed t-test. Error bars represent the standard deviation from 3 biological replicates. The housekeeping gene used in these experiments is tubulin.

For the RNAseq experiment, 5 control and 5 treated samples from a fresh batch of protoplasts were prepared using these conditions, and frozen in liquid nitrogen. 5 rather than the usual 3 replicates were chosen for each treatment following the recommendations of Schurch *et al.* (2016) for the number of samples to be used for RNAseq. Total RNA samples were extracted and quantified on an Agilent 2100 Bioanalyzer; all had a good RNA integrity number > 7.5. This value reflects the ratio of 28S to 18S ribosomal RNA and the degradation of the RNA (Schroeder *et al.*, 2006). Before sending these 10 samples to be sequenced, a qPCR was performed to ensure the treated samples maintained the robust increase in gene expression observed earlier in different samples. This qPCR was performed in a 96-well plate; the amount of noise observed between biological replicates was reduced compared to the results in Figure 5.3, where the qPCR was performed in a 384-well plate. All subsequent qPCRs were therefore carried out in 96-well plates in a QuantStudio3 Real Time PCR System. AT1G56060 was upregulated approximately 5-fold and AT1G72520 was upregulated approximately 15-fold (Figure 5.4). This reinforced the results obtained from the previous qPCRs (Figure 5.3), that chitin treatment was causing differential expression of genes (based on these marker genes), so these 10 samples were sent for RNAseq.

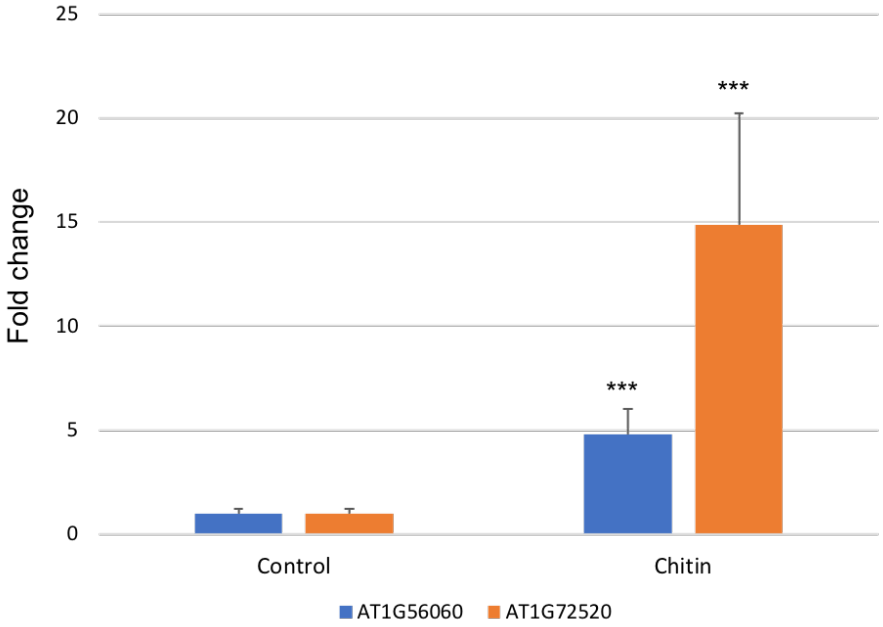


Figure 5.4: Average levels of AT1G56060 and AT1G72520 in control and chitin-treated protoplasts in the samples used for RNAseq. Protoplasts were treated with 10 mg/l chitin or the equivalent volume of buffer (control) for 45 minutes. *** represents $p \leq 0.001$ in a two-tailed t-test. Error bars represent the standard deviation from 5 biological replicates. The housekeeping genes used for normalisation were tubulin and ubiquitin.

5.4.2 Effect of chitin on gene expression in protoplasts compared to gene expression in leaves infected with *B. cinerea*

Following the optimisation process in the previous section, 5 RNA samples from protoplasts treated with 10 mg/l of chitin and 5 RNA samples from protoplasts treated with buffer for 45 minutes for a control were sent for RNAseq. These data are compared to the trends in differential gene expression following *B. cinerea* infection of Arabidopsis to infer whether the chitin-protoplast system makes a good proxy for the *B. cinerea*-leaf system.

75 base pair (bp) paired-end read sequencing was performed on an Illumina HiSeq 4000. Quality control checks were performed on the raw reads with FastQC (Andrews *et al.*, 2010). Reads were confirmed to be of acceptable quality and length (Figure 5.5). Trimming of reads or discarding of low quality samples was not required as the quality score along all positions on the reads was good.

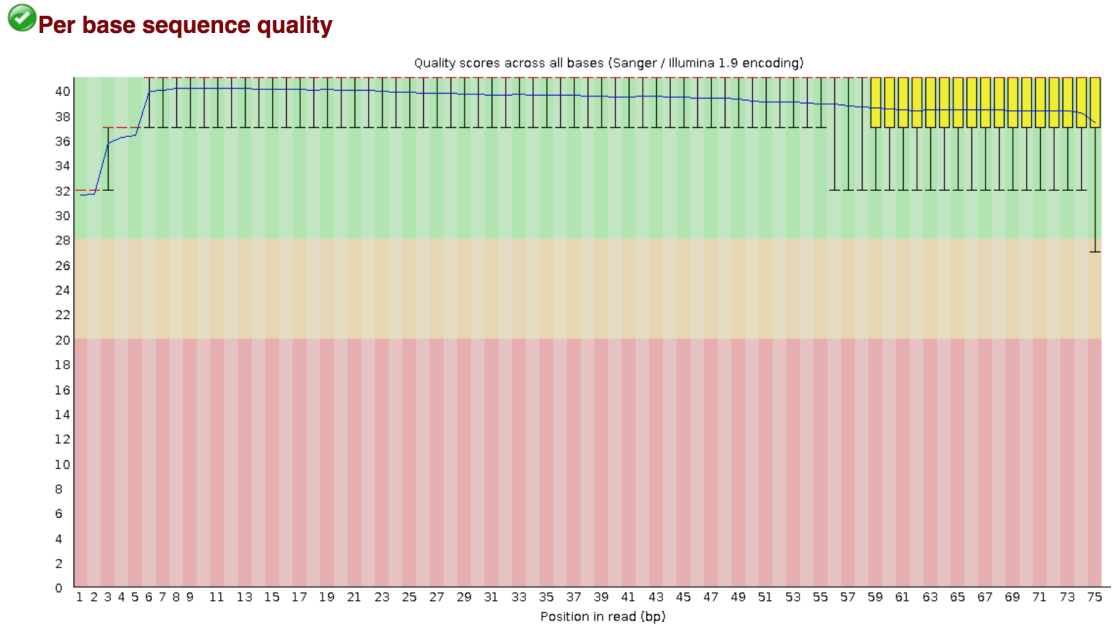


Figure 5.5: A representative FastQC plot of the quality score summarised across read position. Sequencing was carried out using 75 base paired-end reads. The sample in this example is a control sample. In all samples the quality score was above the accepted cut-off of 28, making trimming of reads unnecessary.

These reads were then aligned to an Arabidopsis reference transcriptome with kallisto (Bray *et al.*, 2016). kallisto is a fast RNAseq quantification software that uses pseudo-alignment. Pseudo-alignment refers to the fact that kallisto does not attempt to match reads to their exact origin within transcripts, but just to particular transcripts. This speeds up the process, and it returns output in minutes rather than the hours-long processing of other popular quantification software such as TopHat2 and Cufflinks, without loss in accuracy. k-mers are constructed from a transcriptome de Bruijn graph and the RNAseq reads; a hash approach matches the compatibility of the k-mers of a read with the k-mers of transcripts determines which transcripts it could have originated from.

Altogether, kallisto analysis generated ~356 million mapped reads over the ten samples (Table 5.3). For the main analysis, the splice variants were grouped together by gene using tximport (Soneson *et al.*, 2015), as the analysis did not require quantification of the different transcript isoforms. 37,869 Arabidopsis genes are identified using the latest Araport transcriptome (Cheng *et al.*, 2017). After filtering out genes with fewer than 100 total reads over the 10 samples, 17,826 genes remained. These were analysed with limma-voom (Law *et al.*, 2014) to identify DEGs after treatment. voom is an addition to the differential expression analysis software limma that was initially created for analysing microarray data. limma-voom applies similar normal-based statistical modelling to RNAseq data as the original software. The significant difference in

Sample	Number of reads
Control	27,574,445
Control	36,081,410
Control	33,532,378
Control	42,729,584
Control	26,642,641
Chitin-treated	35,550,907
Chitin-treated	42,881,936
Chitin-treated	41,180,594
Chitin-treated	26,990,984
Chitin-treated	42,874,994

Table 5.3: The number of reads generated across RNAseq samples, as mapped by kallisto.

the two pipelines is voom's incorporation of a mean-variance relationship of reads across a gene in the standard limma empirical Bayes procedure. voom was chosen over other software for differential expression analysis due to its similarity to software for analysing microarray data, making downstream comparisons with microarray datasets fairer.

5.4.2.1 Comparison of chitin treatment with *B. cinerea* infection

3,180 genes were DE due to the chitin treatment relative to the control as determined by limma-voom at the adjusted p-value level of 0.05, which is ∼18% of detected genes. Of these, 1,115 are downregulated and 2,065 upregulated. 117 of 217 possible genes involved in the response to chitin on TAIR (Lamesch *et al.*, 2011) are DE. The most upregulated gene was an F-box protein with unknown function whose expression was 160-fold increased upon treatment. Other highly upregulated genes include a putative receptor, the leucine-rich repeat kinase AT5G39390; WRKY41 which regulates crosstalk between SA and JA pathways and suppresses *PDF1.2* (Higashi *et al.*, 2008); CYP82C3 which is a part of the cytochrome P450 complex, and is potentially involved in the JA response and resistance to *B. cinerea* (Liu *et al.*, 2010). However, the function of most of the other highly upregulated genes has not been studied in detail. The two most downregulated genes are two small auxin-responsive protein, SAUR6 and SAUR25. The response to pathogens represses auxin signalling (Wang *et al.*, 2007), possibly as a way to counter the pathogen's interference of auxin biosynthesis (Glickmann *et al.*, 1998). The top 20 most DEGs are shown in Table 5.4.

Gene Ontology (GO) term enrichment analysis was carried out on the protoplast DEG using the PANTHER Overrepresentation Test (Mi *et al.*, 2013). The number of genes and associated GO terms are given in Dataset H. As expected, response to

Table 5.4: The top 20 upregulated genes as identified by RNAseq, and their differential expression levels, if any, in the other studies. Gene descriptions are obtained from TAIR (Lamesch *et al.*, 2011).

Locus Identifier	FC	Gene Model Description	Gene Name	Notes
AT4G04480	159.56	F-box protein with a domain protein		Upregulated 11-fold in Wan et al., 15-fold in Ramonell et al.
AT1G06137	153.26	transmembrane protein		
AT3G57730	136.40	Protein kinase superfamily protein		
AT4G31950	118.66	member of CYP82C	CYTOCHROME P450, FAMILY 82, SUBFAMILY C, POLYPEPTIDE 3 (CYP82C3)	Upregulated 25-fold in Povero et al., 48-fold in Ramonell et al.
AT4G08950	91.70	Phosphate-responsive 1 family protein	EXORDIUM (EXO)	Upregulated 2-fold in Ramonell et al.
AT4G22030	82.34	F-box protein with a domain protein		
AT1G06135	68.13	transmembrane protein		
AT5G11140	67.60	ARM repeat superfamily protein		
AT4G38420	61.93	SKU5 similar 9	SKU5 SIMILAR 9 (sks9)	
AT4G11070	61.26	member of WRKY Transcription Factor; Group III	(WRKY41)	
AT3G13433	57.40	transmembrane protein		
AT1G51915	56.81	cryptdin protein-like protein		Upregulated 33-fold in Ramonell et al.
AT5G39390	54.83	Leucine-rich repeat protein kinase family protein		
AT2G31345	53.85	transmembrane protein		
AT5G47850	53.69	CRINKLY4 related 4	CRINKLY4 RELATED 4 (CCR4)	Upregulated 140-fold in Ramonell et al.
AT1G51913	53.26	transmembrane protein		
AT1G14550	50.49	Peroxidase superfamily protein		Upregulated 9-fold in Wan et al., 9-fold in Ramonell et al.
AT1G56240	47.51	phloem protein 2-B13	PHLOEM PROTEIN 2-B13 (PP2-B13)	
AT5G67450	43.99	Encodes zinc-finger protein. mRNA levels are elevated in response to low temperature, cold temperatures and high salt. The protein is localized to the nucleus and acts as a transcriptional repressor.	ZINC-FINGER PROTEIN 1 (ZF1)	Upregulated 6-fold in Wan et al., 18-fold in Ramonell et al.
AT3G02840	43.60	ARM repeat superfamily protein		Upregulated 50-fold in Povero et al., 47-fold in Ramonell et al.

chitin is the top overrepresented GO term within the upregulated genes; others include response to stress/stimulus, protein phosphorylation, signal transduction/signalling and protein modification processes. Within the downregulated genes, prominent GO terms include regulation of transcription, regulation of RNA metabolic (and other biosynthetic) processes, and transcription.

In order to establish the overlap between this experiment and published results of Arabidopsis seedlings treated with chitin and its derivatives, the DEGs that were two-fold up- or down-regulated were selected in all the datasets. This resulted in 1,227 protoplast genes upregulated 2-fold and 186 protoplast genes downregulated 2-fold. The overlap between the different experiments was minimal when it came to the downregulated genes (Figure 5.6). 113 genes were upregulated in all 4 experiments. 514 of the genes upregulated in protoplasts were upregulated in at least one other study, while 713 of the upregulated genes were unique to protoplasts treated with chitin. The highest similarity to the protoplast DEG is the Ramonell *et al.* (2005) study which also treated Arabidopsis with chitin.

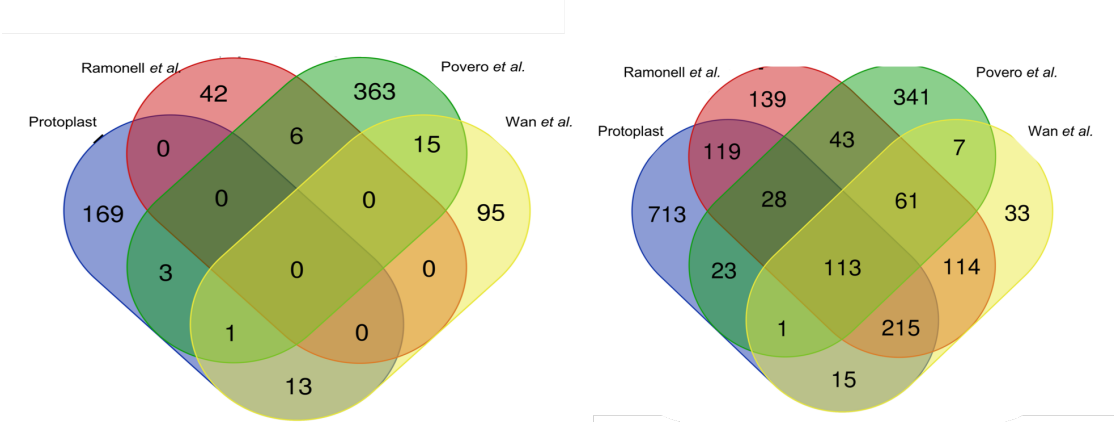


Figure 5.6: Venn diagram of DEGs in protoplasts treated with chitin, and Arabidopsis treated with chitin or its derivatives from Ramonell *et al.* (2005), Povero *et al.* (2011) and Wan *et al.* (2008) studies. Downregulated genes (left) and upregulated genes (right). Only genes that were two-fold up- or downregulated were selected.

In order to find the similarities between the leaf and protoplast response, the list of DEG in *B. cinerea*-infected leaves and the genes in chitin-treated protoplasts were compared. The latter included all genes identified as DE by limma with an adjusted p-value$<$ 0.05, rather than a fold change$>$ 2, as the differential expression of genes to *B. cinerea* infection was measured using microarrays and fold change values are not applicable. Once more, the overlap is smaller between the downregulated genes than between the upregulated genes (Figure 5.7). 752 genes are upregulated in both (36% of the 2065 protoplast DEGs and 18% of the 4112 leaf DEGs) and 259 were downregulated in both (30% of the 1115 protoplast DEGs and only 5% of the 5706 leaf DEGs). Of these DEGs common to both responses, 64 upregulated and 61 downregulated genes

were also present in the *At-Botrytis* network created in Chapter 2, which consists of 883 TFs. This is only 15% of the TFs - a rather small overlap. However, the DEG in leaves are obtained from the list of DEG from each infection time point - 24 in total; the DEG in protoplasts are only obtained from a single time point. Naturally, this would result in fewer observed protoplast DEG. It is also possible that the chosen time point of 45 minutes post protoplast induction does not result in the most number of DEG.

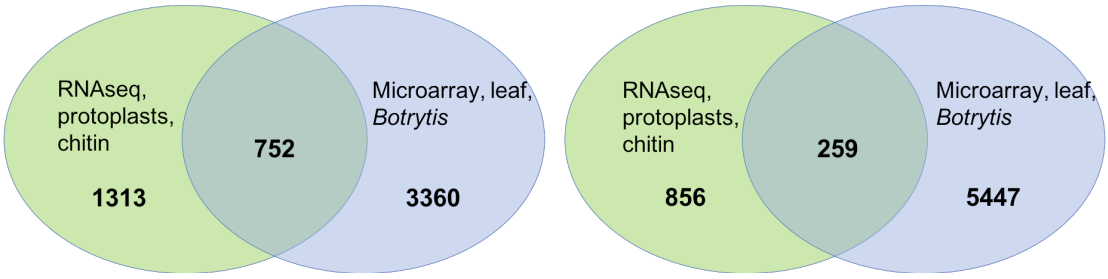


Figure 5.7: Venn diagram of DEGs in protoplasts treated with chitin, and Arabidopsis leaves infected with *B. cinerea*. Gene expression was measured with RNAseq in protoplasts and microarrays in leaves. Downregulated genes (left) and upregulated genes (right).

GO term enrichment analysis was carried out in order to understand the source of this difference between the leaf and the protoplast response. Prominent among the genes downregulated in leaves but not in protoplasts are genes responsible for photosynthesis, a number of biosynthetic processes such as peptide and amide biosynthesis and metabolic processes such as nitrogen, peptide, non-coding RNA, and macromolecule metabolism. The most notable GO terms overrepresented in the 219 genes upregulated in protoplasts and downregulated in leaves infected with *B. cinerea* are the response to bacteria, and protein phosphorylation. GO terms related to phosphorylation and hypoxia are conspicuously absent from the upregulated leaf genes - but are also not downregulated in the leaves, while they are among some of the most overrepresented terms in the upregulated genes of protoplasts. These GO term analyses are included in Dataset H.

Overall, the act of protoplasting, which is stressful to the cells (Marx (2016), Gifford *et al.* (2008), Birnbaum *et al.* (2005)), appears to trigger a number of responses that are not seen in leaves; it may also make the cells less sensitive to chitin, as seen in the comparison with other chitin studies. This is combined with the inability of chitin to trigger a full set of responses like *B. cinerea* which includes the effector-triggered immunity (ETI), particularly in the downregulation of the metabolic genes (Windram *et al.*, 2012). However, the simplest explanation for the disparity in downregulated genes is the fact that the chitin-treated samples were collected after only 45 minutes. This would not allow a majority of mRNAs time to degrade and thereore be downregulated: the mRNA half-life varies from minutes to 26 hours, with an average of 5.9 hours and a median of 3.8 hours (Narsai *et al.*, 2007). The small overlap between the protoplast DEG and the TFs in the *At-Botrytis* network makes simulating Arabidopsis networks for

testing predictions in protoplasts more difficult. Due to the lack of similarity, many genes would have to be left out of the modelling, which would defeat the purpose of a systems-level model; if the genes with dissimilar responses were included in the modelling, this would make simulations inaccurate.

5.4.2.2 Determining similarity of protoplast and leaf systems

In order to determine how similar gene expression is in protoplasts and in leaves, without any treatment, RNAseq data was obtained for 3-week old untreated leaves from Xu *et al.* (2017). This consisted of 2 biological repeats and 27,172 measured genes. This list was compared to the 37,869 Arabidopsis genes identified with the protoplast RNAseq. Genes with no count data in either datasets were eliminated, leaving 15,466 genes. These were then filtered as above to exclude genes with fewer than 70 reads over the 7 samples. limma-voom was used again to find the genes that are differentially expressed between untreated leaves and protoplasts. 8,765 of the 15,466 genes had a fold change of at least 2 (up or downregulated) and an adjusted p-value$<$0.05 between the two conditions. Therefore protoplasting causes more than half of observed genes to differ significantly from leaves. The response to chitin and the response to protoplasting act together in a complex manner, making it hard to decipher the gene response to chitin alone.

However, the leaves used by Xu *et al.* (2017) in their RNAseq experiment were removed from the plant and flash frozen in liquid nitrogen immediately. A better comparison would be with leaves that were treated as in Windram *et al.* (2012): removed from the plant and placed on a tray with agar for at least 24 hours, then flash frozen and used for RNAseq.

5.4.3 Generating rewiring constructs

In order to evaluate the ability of GRN rewiring to enhance the defence response, promoter-gene constructs were cloned and transformed into Arabidopsis plants through floral dipping and protoplasts through polyethylene glycol (PEG)-mediated transformation. The genes and promoters for rewiring are chosen using the transcriptomic time series data available from Windram *et al.* (2012) and known disease phenotypes from literature search. The effects of rewiring are quantified using qPCR in protoplasts, and qPCR and *B. cinerea* infection assays in transformed Arabidopsis plants.

5.4.3.1 Identifying genes suitable for rewiring

The rewirings predicted to have the highest impact on a plant's defence response are the ones which increase the levels of genes that a positive effect on the Arabidopsis defence response, particularly if these genes are downregulated during infection. The downregulation is assumed to be due to the pathogen attempting to subvert the immune response. As explained in Chapter 2, data on genes which show an altered stress response to *B. cinerea* infection when knocked out (KO) or over-expressed (OE) were collected

from published and unpublished sources (Dataset E). 24 genes had a positive effect on defence as demonstrated over multiple experiments, and 9 showed a combination of no phenotype and a detrimental effect when knocked out over multiple experiments. Of the 24 positive regulators of defence, 12 were downregulated in the infection time series. Of these 12, 6 were chosen for being downregulated by a large amount early in the infection process (see Figure 5.8 and Table 5.5). These are *AL1*, *AtWHY2*, *TGA3*, *WRKY3*, *WRKY60*, and *WRKY70*.

The time of differential expression (TOFDE) was determined by performing Student's t-tests at each time point between the control and infected conditions. If $p<0.01$ at two consecutive time points, the first time point was taken to be the TOFDE. The promoters for rewiring were chosen by screening for genes that are highly upregulated early during infection (pre-26 hours post infection) - *ANAC055*, *SAP12* and *ZAT11*. The promoter region included 2000 base pairs of the 5' untranslated region upstream of the starting ATG in an effort to capture as many regulatory interactions as possible.

5.4.3.2 Golden Gate cloning of constructs

In order to test the effect of rewiring Arabidopsis genes, the rewired constructs were made using Golden Gate cloning (Engler *et al.*, 2008). These consisted of the promoter region inserted before the cDNA or coding regions of the chosen genes (see Figure 5.9). The cDNA region consists of the CDS and the 5' and 3' untranslated regions (UTRs). The rewiring was done in col-0 Arabidopsis ecotype, therefore the original gene was still present in the genome and expressed as normal. Protoplast rewiring controls were made by coupling each of these three promoters with *GFP*.

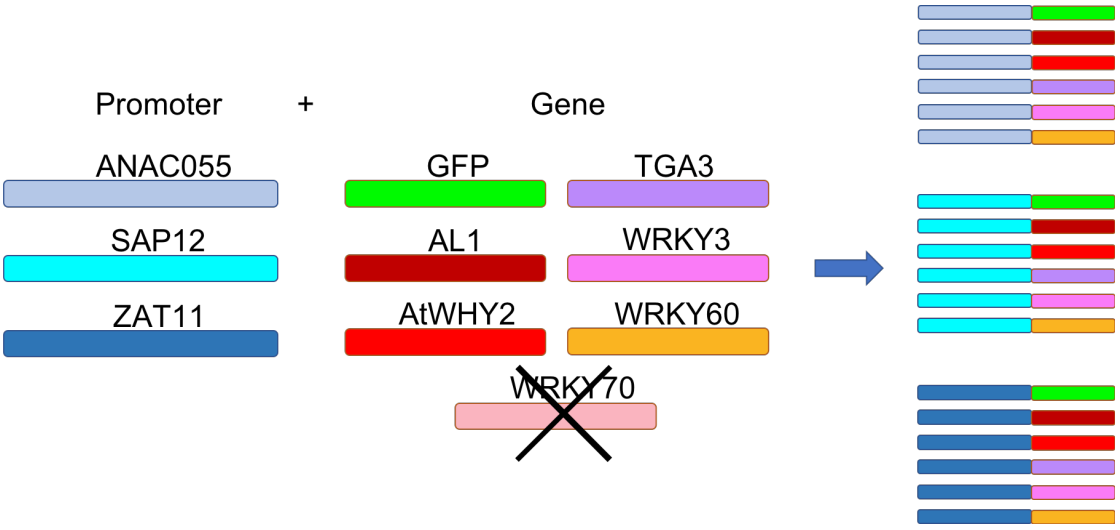


Figure 5.9: Rewired constructs made through Golden Gate cloning. In total, each of 3 promoter regions were combined with each of 6 coding sequences, out of an attempted 7 to create 18 rewiring constructs.

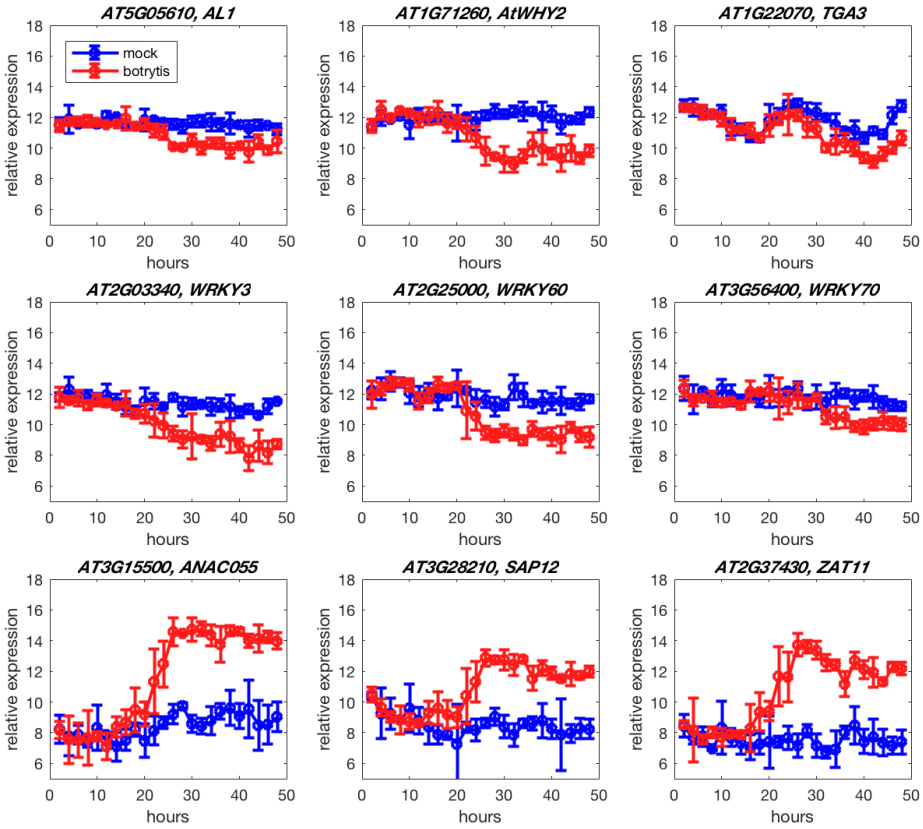


Figure 5.8: Average expression of genes that were selected for rewiring from Windram *et al.* (2012). Constructs were created with the promoters of *ANAC055*, *SAP12*, *ZAT11* with the coding regions of *AL1*, *AtWHY2*, *TGA3*, *WRKY3*, *WRKY60* in order to increase their levels during infection. Lines in blue are expression data from mock treated leaves and lines in red are from *B. cinerea* infected leaves. Error bars represent standard deviation of the 4 biological replicates.

Gene	Locus	Source for positive phenotype
AL1	AT5G05610	Unpublished infection of KO
AtWHY2	AT1G71260	Unpublished infection of KO
TGA3	AT1G22070	Windram et al., 2012
WRKY3	AT2G03340	Lai et al., 2008
WRKY60	AT2G25000	Unpublished infection of KO
WRKY70	AT3G56400	Unpublished infection of KO

Table 5.5: Genes that were selected for rewiring and the source for the phenotype data.

The individual DNA parts were cloned using standards set for plant synthetic biology in Patron *et al.* (2015). A guide to Golden Gate cloning and a plasmid database for this purpose can be found at `http://synbio.tsl.ac.uk/`. These require various parts such as promoters and coding regions to be flanked with specific sequences on either end and be placed in vectors with the appropriate antibiotics. As Golden Gate assembly requires the use of type IIS restriction enzymes *BsaI* and *BpiI*, any recognition sites within the parts themselves needs to be removed or domesticated. This was done through PCR-based mutagenesis to create a synonymous mutation that cannot be targeted by the enzymes. The PCR products were then run on a gel, and the band of the correct size was cut out and purified. Golden Gate Level 0 reactions with *BpiI* were carried out with all the fragments of each part in order to form domesticated parts for *AtWHY2*, *TGA3*, *WRKY3*, *WRKY70*, *ANAC055* and *ZAT11* within the vector pICH41308 (for coding regions) or the vector pICH41295 (for promoter regions). *WRKY60*, *SAP12* and *GFP* did not require domestication and the amplified parts were inserted directly into the appropriate vector. *WRKY70* could not be cloned into its required vector. A Golden Gate Level 1 reaction with *BsaI* was then used to create pairwise promoter-gene constructs of *ANAC055/SAP12/ZAT11* with *AL1/TGA3/WRKY3/WRKY60/GFP* in the vector pICSL86955OD. The *AL1* coding region was not domesticated, as that would have required modification of 5 *BpiI* or *BsaI* sites. Instead, its coding region (CDS rather than cDNA) was inserted straight into a Level 1 construct as it had 3*BpiI* cut sites but no *BsaI* sites. The genes have only a single splice variant, other than AL1 - whose CDS did not vary between the two splice variants. The final vector pICSL86955OD contained, beside a *bar* gene for BASTA resistance and a gene for kanamycin resistance, a 35S terminator of transcription, so that the final constructs were: Promoter - 5' UTR - CDS - 3'UTR - Terminator for all constructs other than those containing *AL1* or Promoter - CDS - Terminator for the *AL1* constructs.

In order to ensure the correct insert was present after each Golden Gate reaction

and that the genes had not acquired point mutations, the plasmids were sent for Sanger sequencing. Primers for sequencing were designed so that they were spaced ~700bp apart (the sequencing returns 850-900bp on average). These confirmed that the gene coding regions had not acquired any mutations through cloning and the promoter regions had at most one mutation.

5.4.4 Effects of re-wiring on the response to chitin in protoplasts

While 18 different constructs were made as described in the section above (Figure 5.9), only some of these were used for transforming protoplasts. The main aim is to directly increase the levels of downregulated genes with positive phenotypes during a triggering of the defence response. These TFs have additional targets downstream so ideally the rewiring would result in the upregulation of other TFs and genes with positive phenotypes without increasing the levels of negative regulators of defence. Therefore in this section and the next, the effect of genes believed to be downstream of the rewiring target is also observed.

Of the 3 promoters, only *SAP12* and *ZAT11* are upregulated in the RNAseq of protoplasts treated with chitin - Figure 5.10. Therefore only these two were chosen for rewiring protoplasts. None of the genes chosen for rewiring were DE in protoplasts treated with chitin - Figure 5.11 - they are only downregulated in infected leaves. However, the rewirings of these genes with *SAP12* and *ZAT11* should still result in a marked increase in observed levels. Additionally, the downregulation in genes in infected leaves is only seen after 24 hpi. Chitin gene expression is only measured at a single time point - 45 minutes after treatment, which may not be enough time for all genes to be DE.

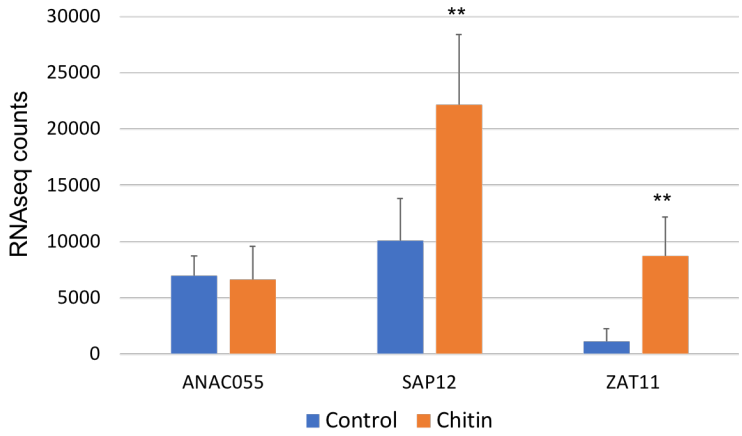


Figure 5.10: RNAseq count data for *ANAC055*, *SAP12* and *ZAT11* from protoplasts treated with and without 10 mg/l chitin for 45 minutes. The generated counts have been scaled for transcript length and library size. Individual bars are the average of 5 biological replicates and error bars are the standard deviation. ** = $p<0.01$ in a two-tailed t-test.

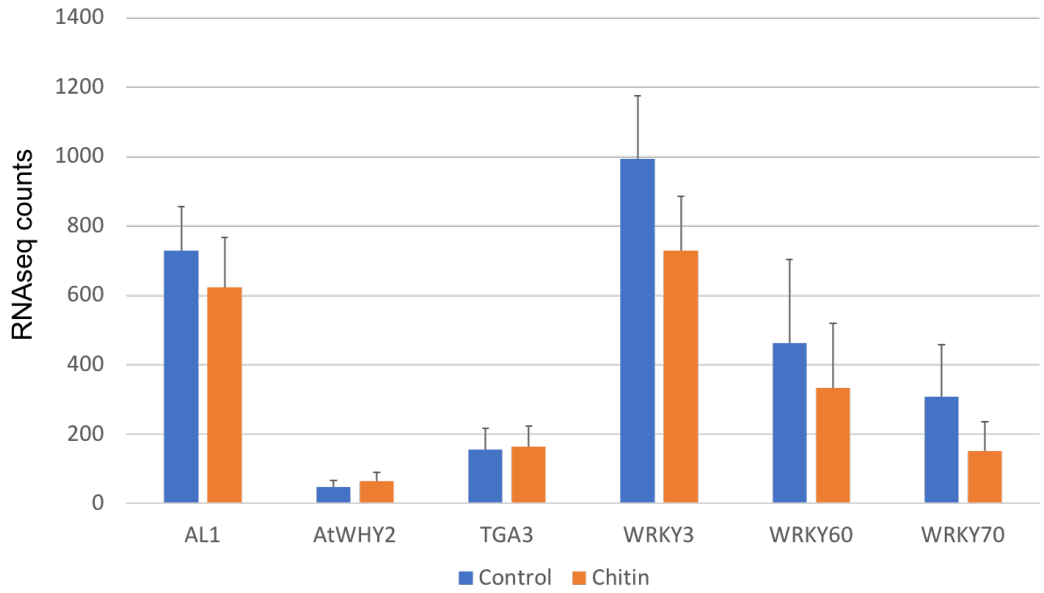


Figure 5.11: RNAseq count data for *AL1*, *AtWHY2*, *TGA3*, *WRKY3*, *WRKY60* and *WRKY70* from protoplasts treated with and without 10 mg/l chitin for 45 minutes. The generated counts have been scaled for transcript length and library size. Individual bars are the average of 5 biological replicates and error bars are the standard deviation. None of the genes are significantly DE.

8 constructs were separately transformed in protoplasts using PEG: *SAP12::AL1*, *SAP12::TGA3*, *SAP12::WRKY3*, *SAP12::GFP*, *ZAT11::AL1*, *ZAT11::TGA3*, *ZAT11::WRKY3*, *ZAT11::GFP*. The constructs were used for transforming Arabidopsis protoplasts that had been extracted using the enzymatic method outlined in Yoo *et al.* (2007). These were left to rest overnight before being induced with 10 mg/l chitin for 45 minutes, as per the RNAseq experiment. RNA was extracted and qPCR of the rewired genes was performed, together with ubiquitin for the housekeeping control. Transformation and treatment of 1 biological replicate was performed on one day and the process was repeated for another 2 biological replicates on a different day.

5.4.4.1 Effect on rewired genes

In all cases, the rewired gene was highly upregulated, as expected, when comparing its levels in the *SAP12::GFP* or *ZAT11::GFP* control to the rewired counterpart (Figure 5.12). The average expression of *AL1* (relative to ubiquitin) in *SAP12::AL1*-transformed protoplasts is 209, whereas in the *SAP12::GFP* control it is 0.4, for instance. This is a change of 560-fold. *TGA3* is increased 523-fold in the *SAP12::TGA3* protoplasts and *WRKY3* is increased 535-fold in the *SAP12::WRKY3* protoplasts. For the *ZAT11* rewirings, *AL1* increased 220-fold, *TGA3* increased 287-fold and *WRKY3* increased 109-fold compared to their expression in the *ZAT11::GFP* control.

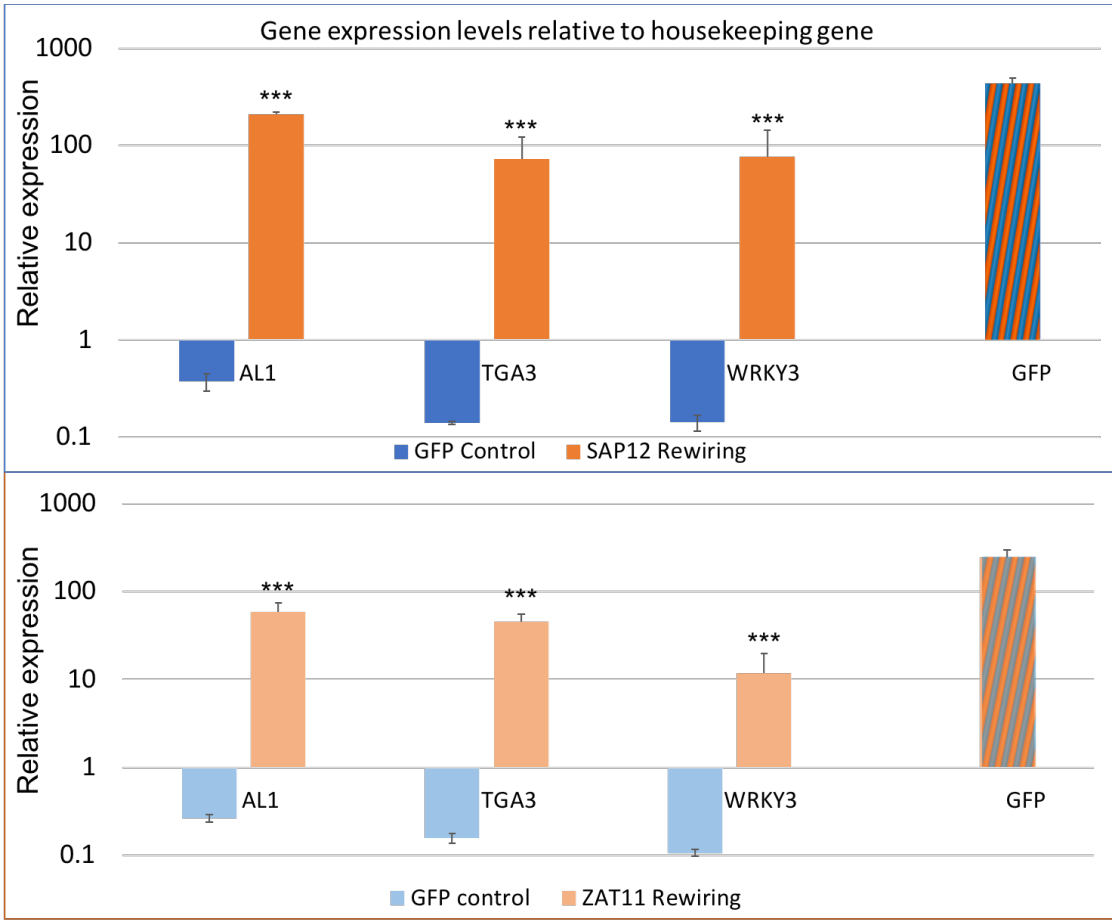


Figure 5.12: Average expression levels of *AL1*, *GFP*, *TGA3* and *WRKY3* relative to the housekeeping gene ubiquitin in protoplasts transformed with rewired constructs and **all** treated with 10 mg/l chitin for 45 minutes. Top: levels of *AL1/TGA3/WRKY3* in the *SAP12::GFP* protoplasts are shown in dark blue and the levels of *AL1/TGA3/WRKY3* in *SAP12::AL1*, *SAP12::TGA3*, and *SAP12::WRKY3* protoplasts, respectively, are shown in dark orange. There is no control for the levels of *GFP* in the *SAP12::GFP* protoplasts. Bottom: levels of *AL1/TGA3/WRKY3* in the *ZAT11::GFP* protoplasts are shown in light blue and the levels of *AL1/TGA3/WRKY3* in *ZAT11::AL1*, *ZAT11::TGA3*, and *ZAT11::WRKY3* protoplasts, respectively, are shown in light orange. There is no control for the levels of *GFP* in the *ZAT11::GFP* protoplasts. Relative expression is plotted on a **log scale** in order to be able to show the values in the GFP controls and the rewired samples on the same plot. Individual bars represent the average of 3 biological replicates analysed on different days and bars represent the standard deviation. *** = $p<0.001$ in a two-tailed t-test.

Interestingly, although all four coding regions (*AL1*, *GFP*, *TGA3* and *WRKY3*) had the same promoter region (either *SAP12* or *ZAT11*), they showed different expression

levels after rewiring. For instance, *GFP* under both *SAP12* and *ZAT11* had the highest expression value and is probably the most representative of the promoter's strength due to the lack of crosstalk or feedback of *GFP* downstream. This echoes the results of Isalan *et al.* (2008) and Windram *et al.* (2017) that the coding region and the regulatory region both play a part in determining the final gene expression, with additional complexity arising from the interplay between the gene AND the regulatory region. AL1, TGA3 and WRKY3 regulate downstream genes and thus directly or indirectly affect the binding of regulators to the *SAP12*/*ZAT11* promoters, effectively providing a negative feedback loop. For instance, by looking at the consensus *At-Botrytis* network with 883 nodes and 8830 edges, there are a number of pathways leading from AL1/TGA3/WRKY3 to regulate *SAP12*/*ZAT11*. AL1 is predicted to regulate *ZAT11* via AT4G28640, otherwise known as IAA11, which is a repressor of early auxin response genes. AL1 is also predicted to regulate *SAP12* via 4 pathways which consist of 2 intermediary transcription factors. TGA3 is also predicted to regulate *SAP12*/*ZAT11* via numerous pathways that include 3 intermediary TFs. There is only one pathway linking WRKY3 to *ZAT11* via 2 intermediary TFs (via ORC1A and the bHLH AT5G56960) and numerous pathways linking it to *SAP12* via 3 intermediary TFs. It appears that in cases where there are many pathways linking the rewired TF to its rewired promoter, the negative feedback is less pronounced. For instance, there is only one pathway linking AL1 and WRKY3 to the *ZAT11* promoter, but several pathways linking AL1 and WRKY3 to the *SAP12* promoter. In the former case, the expression of *AL1* and *WRKY3* in their rewired states is much more reduced than expected. It is possible that having multiple pathways leading to a promoter will result in the integration of multiple repressive and activating signals and a less pronounced result. However, given that a transformation control was not included in the experiments, this variation in gene expression level of the rewired genes under the same promoter could simply be due to different transformation efficiencies of the replicates.

5.4.4.2 Effect on genes downstream of the rewired genes

In addition to measuring the levels of the rewired gene, qPCR analysis of marker genes and genes downstream of the rewired gene was carried out, in order to have a better understanding of the potential impact on disease resistance. Genes such as *PAD3*, *PDF1.2*, *ERF1* and *WRKY33* have been implicated in the response to disease, as elaborated below, and their levels may give an indication of how this response is faring. In addition to these known defence genes, a few TFs downstream of the rewired genes in the *At-Botrytis* network were also chosen for qPCR analysis.

PDF1.2 is an ET- and JA-responsive antifungal defensin (Penninckx *et al.*, 1998) and is commonly used as a marker for the activation of these two pathways. *WRKY33* is another positive regulator of defence against necrotrophs and a negative regulator of defence against *P. syringae* (Zheng *et al.*, 2006). *WRKY33* appears to regulate the SA-related host response by repressing it and the JA-related response by downregulating JAZ proteins (Birkenbihl *et al.*, 2012), and also repressing the ABA response by downregulating the ABA biosynthetic genes *NCED3* and *NCED5* (Liu *et al.*, 2015). *ERF1*

activates the expression of *PDF1.2* (Pré *et al.*, 2008). Similarly to *WRKY33*, *ERF1* overexpression acts positively in the response to necrotrophs and negatively in the response to biotrophs such as *P. syringae* (Berrocal-Lobo *et al.*, 2002). It is thought to act as a mediator between EIN3, and downstream ethylene-responsive genes (Solano *et al.*, 1998) and is also activated by JA signalling (Lorenzo *et al.*, 2003). A promising rewiring would increase the levels of these genes.

In the RNAseq experiment measuring the impact of chitin on protoplast gene expression, *ERF1* was significantly downregulated, *WRKY33* was significantly upregulated and *PDF1.2* was upregulated, but not significantly so (Figure 5.13).

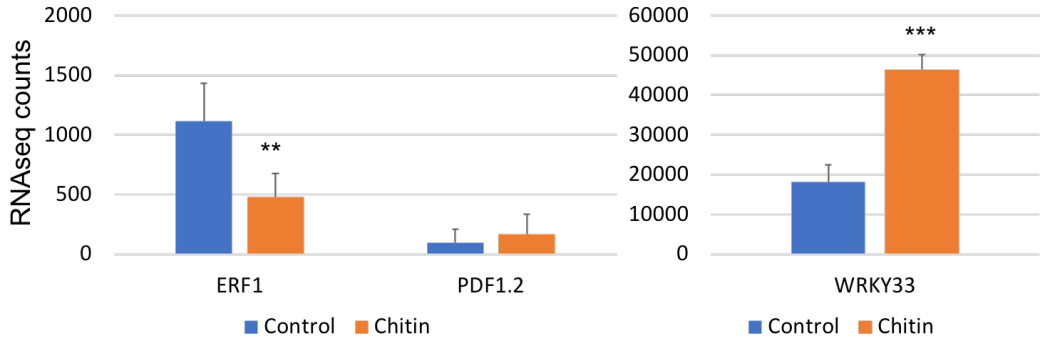


Figure 5.13: RNAseq count data for *ERF1*, *PDF1.2*, and *WRKY33* from protoplasts treated with and without 10mg/l chitin. The generated counts have been scaled for transcript length and library size. Individual bars are the average of 5 biological replicates and error bars are the standard deviation. ** = $p<0.01$ and *** = $p<0.001$.

Of the 3 defence markers, only *WRKY33* shows significant differential expression upon rewiring. It was downregulated in two cases, in the *SAP12::WRKY3* and *ZAT11::TGA3*-transformed protoplasts. As WRKY33 is a positive regulator of defence, these rewirings may be detrimental to the plant. The other 2 defence markers, *PDF1.2* and *ERF1* did not show any difference in expression (see Figure 5.14).

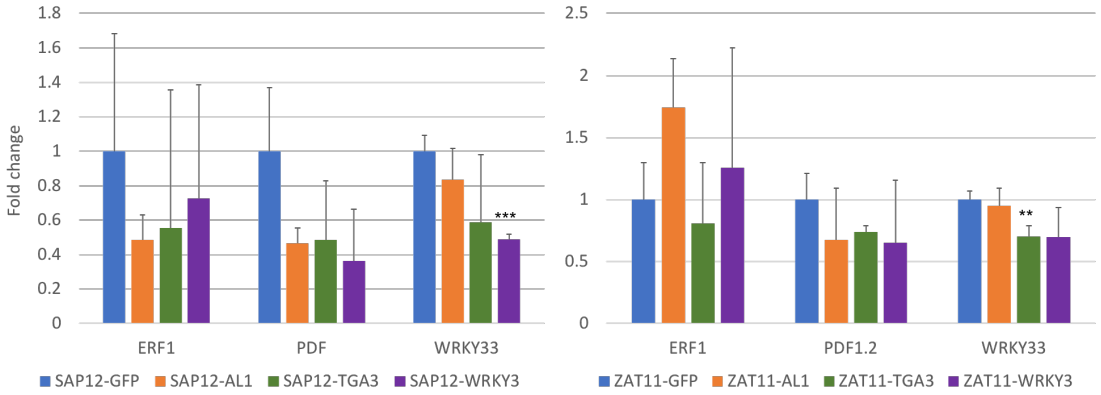
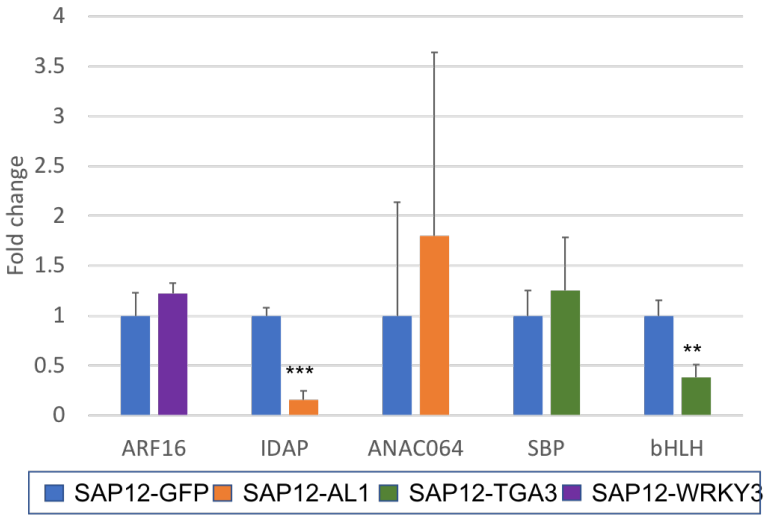


Figure 5.14: Expression fold change of known defence genes in protoplasts rewired with the promoter of *SAP12* or *ZAT11* and treated with chitin, measured with qPCR. The levels of *ERF1, PDF1.2, WRKY33* were measured using qPCR in the plasmids rewired using the *SAP12* promoter and normalised so that their expression levels in the contol *SAP12::GFP* was 1, and the other levels were relative to it (left). The levels of *ERF1, PDF1.2, WRKY33* were measured using qPCR in the plasmids rewired using the *ZAT11* promoter and normalised so that their expression levels in the contol *ZAT11::GFP* was 1, and the other levels were relative to it. Individual bars represent the average of 3 biological replicates performed on different days and bars represent the standard deviation. ** = $p<0.01$ and *** = $p<0.001$ in a two-tailed Student's t-test.

The other genes chosen for gene expression analysis via qPCR were either downstream of the rewired gene in the *At-Botrytis* network, differentially expressed in a KO experiment of the rewired gene, or both. In addition, TFs with a verified role in defence against *B. cinerea* were prioritised, as were TFs belonging to TF families involved in the stress response, such as NACs and WRKYs. Details of these reporters can be found in the tables in Figure 5.15 and 5.16. 4 out of the 11 tested reporters were differentially expressed relative to their expression in the *SAP12::GFP* or *ZAT11::GFP* controls. These are: *AT4G31270*, *IDAP2* in *SAP12::AL1*; *AT5G57150*, a bHLH TF in *SAP12::TGA3*; *AT2G22300*, *CAMTA3* in *ZAT11::WRKY3*; and *AT1G59750*, *ARF1* in the *ZAT11::AL1* protoplasts. It is surprising that *SBP* in the *SAP12::TGA3* protoplasts and *WRKY70, WRKY8* in the *ZAT11::TGA3* protoplasts were not significantly different to the control, as they have been found to be differentially expressed in *tga3* KO lines (Windram *et al.*, 2012). It was expected that their levels would change, given the 523- and 287-fold increase of *TGA3* in the *SAP12* and *ZAT11* rewirings, respectively. However, out of the 10 genes chosen for being downstream of the rewired gene in the *At-Botrytis* network, 4 were differentially expressed in rewired protoplasts. This highlights the usefulness of having a network such as the *At-Botrytis* created for informing experiments by narrowing down the choice of candidates - even though the network was made from data gathered in slightly different circumstances.

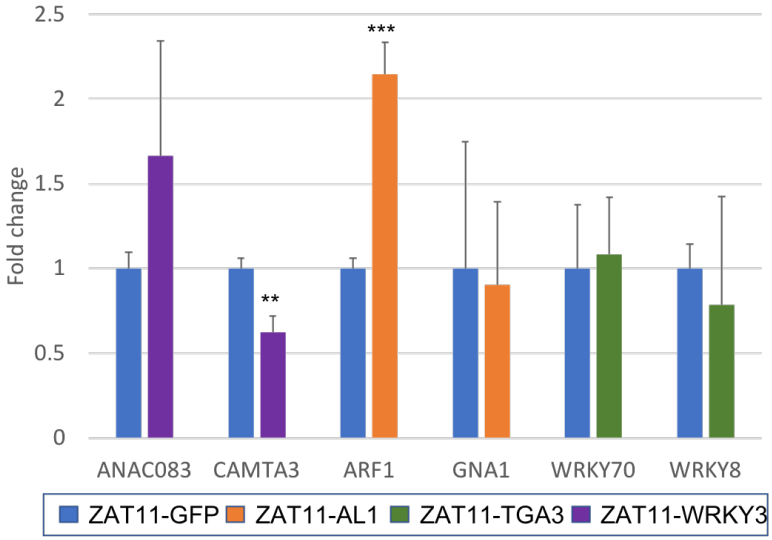
While the rewired gene show an increase 100-560 fold, none of the measured

downstream gene expressions have as marked a change. The largest observed change is in *IDAP2*, which decreases 6-fold in the *SAP12-AL1* construct. IDAP2 is a regulator of DNA methylation and prevents epigenetic silencing (Duan *et al.*, 2017). The calmodulin-binding transcription activator CAMTA3, which also showed downregulation in the *ZAT11-WRKY3* rewiring, has been shown to suppress the defence response to both *P. syringae* and *B. cinerea*, possibly by repressing *WRKY33* expression (Galon *et al.*, 2008). ARF1 is an auxin response factor which plays a positive role in defence (unpublished PRESTA data) and represses auxin-induced genes (Ellis *et al.*, 2005). The function of bHLH protein AT5G57150 has so far been unstudied. Given the decrease of CAMTA3 (a negative regulator of defence) and the increase of *ARF1* (a positive regulator of defence), the most promising rewirings to proceed with based on these results are *ZAT11::WRKY3* and *ZAT11::AL1*.



Reporter AGI	Reporter name	Rewired gene	Source	Result
AT4G31270	*IDAP2*	*AL1*	*At-Botrytis* network	Downregulated
AT3G56530	*ANAC064*	*AL1*	*At-Botrytis* network	Not DE
AT5G57150	*bHLH*	*TGA3*	*At-Botrytis* network + DE in KO	Downregulated
AT1G76580	*SBP*	*TGA3*	*At-Botrytis* network + DE in KO	Not DE
AT4G30080	*ARF16*	*WRKY3*	*At-Botrytis* network	Not DE

Figure 5.15: Expression fold change of reporter genes for protoplasts rewired with the promoter of *SAP12* and treated with chitin, measured with qPCR. Genes along the x axis were selected for being downstream of the rewired gene and their levels measured using qPCR in 3 biological replicates in rewired protoplasts and the *SAP12::GFP* control. The expressions have all been normalised on a gene by gene basis, by setting that particular gene expression in the control *SAP12::GFP* protoplasts to 1. Individual bars represent the average of 3 biological replicates performed on different days and bars represent the standard deviation. ** = $p<0.01$ and *** = $p<0.001$ in a two-tailed Student's t-test. The results are also summarised in the table below the plot. The reporter genes, their AGI, the rewired gene they are downstream of are all given. Also noted is the source for believing the marker gene is downstream of the rewired gene, and whether or not they are DE in the rewiring.



Reporter AGI	Reporter name	Rewired gene	Source	Result	Phenotype
AT1G59750	*ARF1*	*AL1*	*At-Botrytis* network	Upregulated	Positive
AT5G15770	*GNA1*	*AL1*	*At-Botrytis* network	Not DE	Not tested
AT3G56400	*WRKY70*	*TGA3*	*At-Botrytis* network + DE in KO	Not DE	Positive
AT5G46350	*WRKY8*	*TGA3*	DE in KO	Not DE	Not tested
AT5G13180	*ANAC083*	*WRKY3*	*At-Botrytis* network	Not DE	Negative
AT2G22300	*CAMTA3*	*WRKY3*	*At-Botrytis* network	Downregulated	Negative

Figure 5.16: Expression fold change of reporter genes for protoplasts rewired with the promoter of *ZAT11* and treated with chitin, measured with qPCR. Genes along the x axis were selected for being downstream of the rewired gene and their levels measured using qPCR in 3 biological replicates in rewired protoplasts and the *ZAT11::GFP* control. The expressions have all been normalised on a gene by gene basis, by setting that particular gene expression in the control *ZAT11::GFP* protoplasts to 1. Individual bars represent the average of 3 biological performed analysed on different days and bars represent the standard deviation. ** = $p<0.01$ and *** = $p<0.001$ in a two-tailed Student's t-test. The results are also summarised in the table below the plot. The reporter genes, their AGI, the rewired gene they are downstream of are all given. Also noted is the source for believing the marker gene is downstream of the rewired gene, and whether or not they are DE in the rewiring, and whether they play a role in defence against *B. cinerea*.

5.4.5 Effects of stable re-wiring of Arabidopsis leaves on *B. cinerea* infection

Finally, some rewiring constructs were introduced into Arabidopsis plants using floral dipping. The next generation (the T1 plants) were then assessed using an infection assay to determine whether the rewiring had an effect on the resistance to *B. cinerea*. The effects of rewiring Arabidopsis plants are also compared to the results in protoplasts

Table 5.6: Rewirings used in transforming Arabidopsis through floral dippings. Promoters are in the columns and genes are in the rows, these form promoter::gene constructs, such as ANAC055::AL1.

	ANAC055	SAP12	ZAT11
AL1	✓	✗	✓
AtWHY2	✓	✓	✓
TGA3	✓	✓	✓
WRKY3	✓	✓	✓
WRKY60	✓	✓	✓
GFP	✗	✗	✗

presented in the previous section.

14 Golden Gate-assembled constructs in plasmid pICSL86955OD were introduced into Arabidopsis by floral dipping (Clough and Bent, 1998). These are *ANAC055-AL1/AtWHY2/TGA3/WRKY3/WRKY60*, *SAP12-AtWHY2/TGA3/WRKY3/WRKY60* and *ZAT11-AL1/AtWHY2/TGA3/WRKY3/WRKY60* (see Table 5.6). Two separate dippings were performed for each of 14 constructs and kept separate to determine whether the phenotypes could be replicated. This is necessary because the phenotype could be due to the insertion position of the rewiring construct, rather than the rewiring itself. Observing the same *B. cinerea*-susceptibility phenotype in two dipping replicates would strengthen our belief in the rewiring as the cause of the change.

T1 seeds from the transformed Arabidopsis were then sown on BASTA-soaked trays for selection. 10 plants for each of the two dippings for *ANAC055-AL1/AtWHY2*, *SAP12-AtWHY2* and *ZAT11-AL1/AtWHY2* that grew on BASTA were selected and transferred to normal soil. These were grown to flowering and the seeds produced (the T2 generation) collected. These rewirings were not tested for their impact on the *B. cinerea* infection.

10 plants for each of the two dippings for *ANAC055-TGA3/WRKY3/WRKY60* and *SAP12-TGA3/WRKY3/WRKY60* that grew on BASTA were selected and transferred to normal soil. The infection assay was carried out on these plants. While carrying out the assay on the T1 generation is not ideal, due to the plants being heterozygous (or potentially having multiple insertion sites in the genome), the effects of rewiring should still be noticeable with a single copy in the genome. The plants were grown for 5 weeks, after which 3 leaves were picked from each plant (for a total of 30 leaves for each rewiring construct) to infect with *B. cinerea* and measure lesion size over the following days to assess disease resistance. An additional leaf from each plant was infected with *B. cinerea* and frozen in liquid nitrogen after 28 hours for analysis of gene expression with qPCR. This time point was chosen as it represents the peak in upregulation of *ANAC055*, and

SAP12 (see Figure 5.8). Arabidopsis transformed with a pEarleyGate201 plasmid with *35S::GFP* and the *bar* gene were subject to the same treatment as the rewired T1 plants. Together with Col-0 Arabidopsis grown concurrently on normal soil, these were used as controls. The *35S::GFP* line was used as a control together with the WT Col-0 in case the period spent growing on BASTA affected the disease response.

4 infected leaves were chosen from 4 individual plants rewired with the same construct for qPCR assays. The levels of the rewired gene was measured in each, as well as the levels of *PDF1.2* and *WRKY33* in the *SAP12::WRKY3* plants. As expected from the microarray data, rewiring with the *ANAC055* promoter resulted in higher gene expression levels of the rewired gene than with the *SAP12* promoter - Figure 5.17. All rewirings showed significant but noisy upregulation of the target gene, other than *SAP12::WRKY60*, which was upregulated on average, but not significantly so. The upregulation ranged from 2 to 20-fold, which is lower than that observed in protoplasts. However, PEG transformation of protoplasts leads to multiple plasmids being uptaken by the protoplasts, which would naturally result in more copies of the rewired gene being made.

When looking at average gene expressions (relative to ubiquitin), *ANAC055* and *SAP12* rewirings resulted in similar final levels of *TGA3* and *WRKY3* (Figure 5.17, bottom). This was around 0.2-0.3 for the genes when they were rewired with *ANAC055* and 0.1 when they were rewired with *SAP12*. In protoplasts rewired with *SAP12*, *TGA3* and *WRKY3* also had similar levels (Figure 5.12). However, the expression of *WRKY60* upon rewiring was much smaller - 0.008 on average with both promoters. This, once more, echoes the results of Isalan *et al.* (2008) and Windram *et al.* (2017) that the coding region and the regulatory region both influence the final gene expression. In addition, two defence markers, *WRKY33* and *PDF1.2*, were also measured in the *SAP12::WRKY3* rewired leaves. Neither of these showed any significant change compared to the control, although the levels of *PDF1.2* were decreased on average.

Infected leaves were photographed at 3 time points - 48, 62 and 72 hours post infection and the lesion size on each leaf was measured with Fiji (Schindelin *et al.*, 2012). Lesion size for the various T1 plants and the *GFP* and Col-0 controls is shown in Figures 5.18 and 5.19. Of all the assessed rewirings, only one resulted in a reduction of lesion size: the *SAP12::WRKY3* insertions, albeit only at 48 and 72 hours post infection. This is surprising, given that this construct decreased the levels of the *WRKY33* defence gene in protoplasts. However, qPCR performed on the rewired leaves show that the levels of *WRKY33* are not affected. The *ANAC055::TGA3* plants also seemed to be slightly more susceptible to disease (Figure 5.19). One dipping of *ANAC055:WRKY3* also had significantly smaller lesion size at 48 and 72hpi, although this was not seen with the other dipping. This highlights one of the problems with using multiple T1 generation plants - the measured lesion sizes are vary widely, which could be due to multiple inserts in a plant disrupting important genes. However, observing a phenotype even under such noisy conditions - like that seen in the *SAP12::WRKY3* rewired plants, means that the rewiring is very effective in enhancing the defence response.

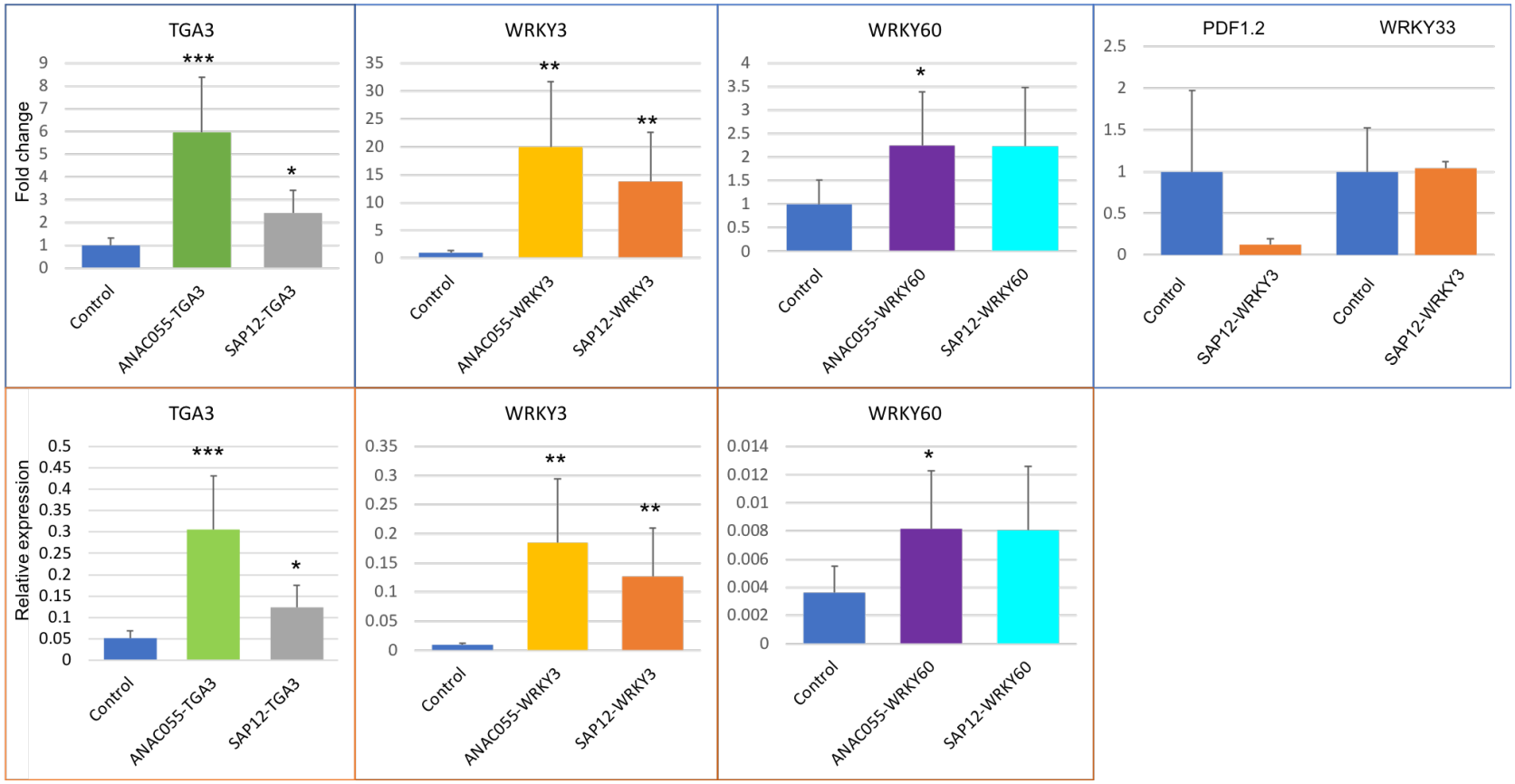


Figure 5.17: Gene expression changes following rewiring in Arabidopsis leaves and *B. cinerea* infection. Gene expression was measured in leaves infected with *B. cinerea* for 28 hours. The control group consists of the average expression from 3 col-0 leaves and 3 leaves with *35S::GFP*. The rewired expressions are the average of leaves from 4 different plants. Fold change in average gene expression in rewired leaves compared to the Col-0 and *35S::GFP* controls (top). The expression of the rewired genes have all been normalised respective to their controls so that the gene expression in the controls is set to 1 (top). Average gene expression following rewiring and relative to the housekeeping gene ubiquitin (bottom), measured by qPCR. Error bars represent standard deviation. * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$ in a two-tailed t-test.

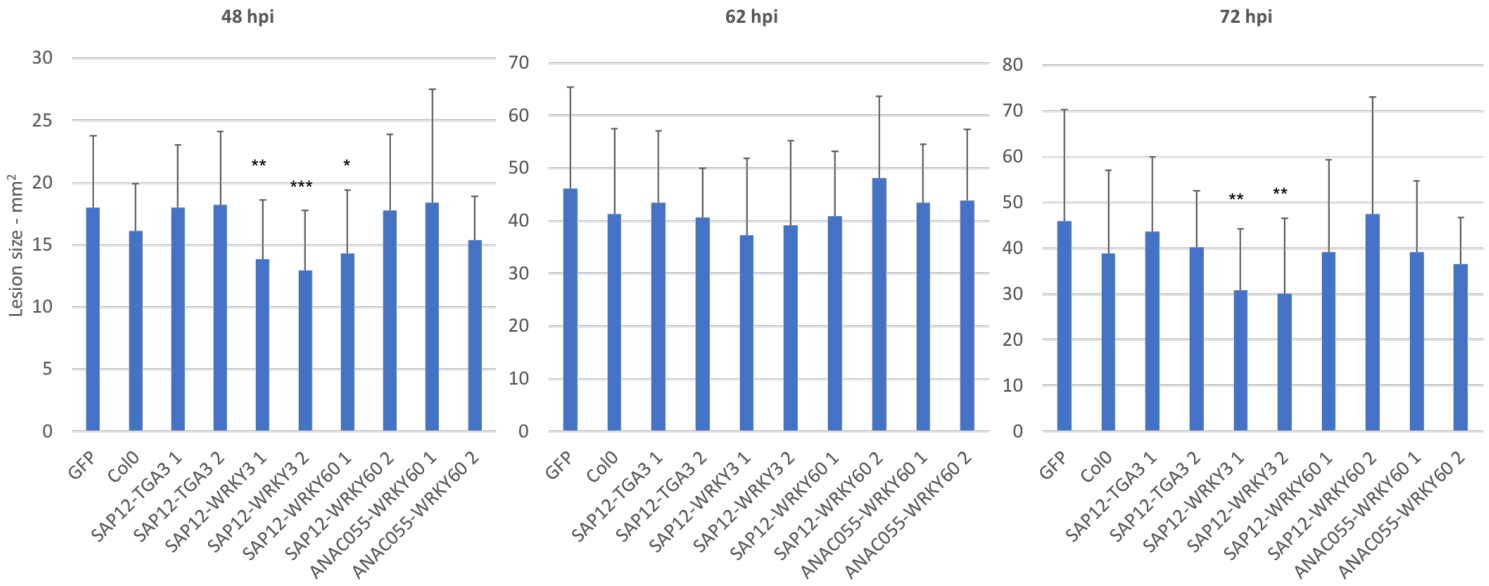


Figure 5.18: Mean lesion size of rewired and control leaves infected with *B. cinerea* at 48, 62 and 72 hpi. 30 leaves were used for each sample. 'SAP12-WRKY3 1' and 'SAP12-WRKY3 2' refer to dippings 1 and 2 of that construct. Error bars represent standard deviation. * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$ in a two-tailed t-test with the *GFP* and Col-0 measurements as the control.

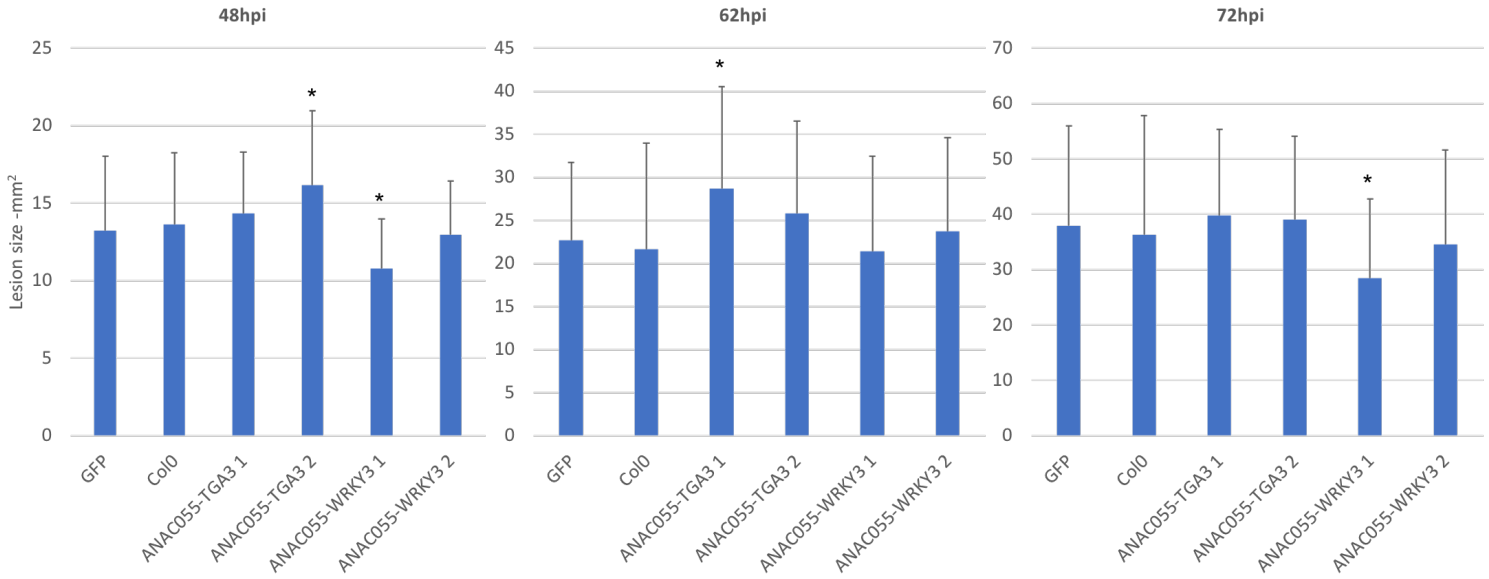


Figure 5.19: Mean lesion size of rewired and control leaves infected with *B. cinerea* at 48, 62 and 72 hpi. 30 leaves were used for each sample. 'ANAC055-WRKY3 1' and 'ANAC055-WRKY3 2' refer to dippings 1 and 2 of that construct. Error bars represent standard deviation. * = $p<0.05$ in a two-tailed t-test with the *GFP* and Col-0 measurements as the control.

5.5 Discussion

This chapter explores the possibility of using protoplasts as a system for understanding and manipulating the Arabidopsis defence response to the necrotroph *B. cinerea*, a pathogen of huge economic importance. As an infection assay system cannot be used with protoplasts, a MAMP present in fungal cell walls, chitin, was used to elicit the defence response instead. Therefore, two comparisons were made - whether chitin is a good proxy for *B. cinerea* and whether gene expression in protoplasts and in Arabidopsis leaves are sufficiently similar.

Using protoplasts for testing rewiring constructs presents many advantages. Protoplasts are straightforward to extract, and to transform. They can be used on a high throughput scale - transformations and inductions can be carried out concurrently in 6-, 24- or 96-well plates. Additionally, chitin elicits the defence response in protoplasts after only 45 minutes of induction. While chitin and *B. cinerea* share a number of gene responses (Figure 5.2), not much is known about using chitin as a proxy for infection in protoplasts. For this reason, a novel RNAseq dataset was generated to determine the extent of this shared response and the comparisons that can be safely drawn between the network response in protoplasts and in leaves. Around 10% of DEGs during infection were differentially expressed in the same way in protoplasts treated with chitin, and this made up 30% of all the differentially expressed protoplast genes. Additional analysis could be done in the future, such as quantifying differential expression of alternatively spliced transcripts, rather than at the gene level (Marquez *et al.*, 2012). However, the short induction time does not leave enough room for many genes to appear downregulated as the available pre-induction mRNA takes longer than 45 minutes to degrade. This was seen in the small overlap between downregulated genes in protoplasts induced with chitin and leaves infected with the necrotroph (Figure 5.7).

A major reason for the inconsistency in these results is the difference in background gene expression - in untreated Col-0 leaves and protoplasts, more than half of the observed gene expressions are significantly different. Hundreds of DEG found in leaves induced with chitin or its derivatives were not DE in the protoplasts. The act of protoplasting itself causes stress genes to be up- or downregulated, which may then not respond to the addition of chitin. A comparison of data between leaves and protoplasts from plants of the same age showed that nearly 8,800 genes of 15,500 had a 2-fold or more change. Therefore the protoplast system should be used with caution, by perhaps restricting studies to genes known to act similarly in both systems. A fairer comparison between the protoplast and *B. cinerea*-infected leaves would have been to only look at the early DE TFs from the leaf time series, thereby ignoring the changes caused by the activation of the ETI. The earliest changes were observed around 18-20 hours post infection and these would consist the basal immune response which is also induced by chitin. However, the overlap between the DEGs generated from both datasets was so small that it was thought prudent to include all the DEGs from the time series for the comparison.

Once the characterisation of protoplasts was carried out, enhancement of the de-

fence response was attempted by increasing the levels of positive regulators of defence through their rewiring with promoters of early responders among highly upregulated genes. Protoplasts combined with qPCR proved to be a useful medium for exploring the relationships between different genes. However, the results obtained with protoplasts may be an extreme version of events, due to the tendency of protoplasts to get transformed with multiple copies of a plasmid. According to the protoplast qPCR results, *ZAT11-AL1* and *ZAT11-WRKY3* show promise as they upregulate a positive regulator of defence and downregulate a negative regulator of defence, respectively. A challenge with using protoplasts is the lack of an obvious phenotype to indicate an enhanced defence response. Therefore 3 marker genes, with well known defence roles - *ERF1, PDF1.2, WRKY33* - were measured to determine whether their levels would give a good indication of the disease progression in plants. However, their levels did not match with the observed disease phenotypes in rewired Arabidopsis leaves - the levels of *WRKY33* were downregulated in *SAP12::WRKY3* protoplasts (Figure 5.13), but not in *SAP12::WRKY3* leaves (Figure 5.17) and furthermore, *SAP12::WRKY3* plants shows reduced lesion size. If continuing with the protoplast system, more marker genes would need to be tested. An alternative to using qPCR would be to transform protoplasts with the rewired construct and a separate construct containing the promoter of a reporter gene fused with luciferase. This would allow easy reporter gene detection with a luciferase assay. With a pipetting robot, the whole process of transforming protoplasts and luciferase screening could then be easily automated.

Infection assays were carried out in 6 transgenic rewired T1 plants. Being the first generation after floral dipping, they are heterozygous for the rewiring. Additionally, they may have one or more inserts in their genome. This is usually avoided by choosing T2 plants where the ratio of BASTA resistant : BASTA susceptible plants is 3:1 (meaning that its T1 parent is heterozygous with a single insertion site). Homozygous T3 plants from the T2 parents are then usually used for assays to determine the effect of the genetic engineering. Due to time restraints, T1 plants were used, and only 6 rewirings were grown and infected with *B. cinerea*. In order to overcome the issues with using plants from the T1 generation, 3 leaves from 10 different plants from each dipping were used for the infection assay. This meant that if a true decrease or increase in lesion size was observed, it would be robust across a range of insert positions and number of inserts. While all rewirings resulted in increased expression of the rewired gene, this fold change was only 2-20 fold, which was in addition to the native gene (this was not knocked out). Genomic integration at a single site naturally reduces the magnitude of rewired gene expression compared to the multiple copies achieved through transformation with plasmids. Unfortunately this could not be compared to the expression levels of *ANAC055* and *SAP12* as their levels were not measured with qPCR. For instance, it would be interesting to see how the strength of the *ANAC055/SAP12* promoter compares when used for regulating *ANAC055/SAP12* and when used for regulating the rewired genes. Other than downstream interference due to the rewired gene, it is possible that not all the regulatory inputs were captured using a promoter region of only 2000bp. There may also be an impact on the levels of *ANAC055/SAP12* if they have to 'share' regulators

with the rewired gene.

The only plants with a robust phenotype were the ones carrying a *SAP12::WRKY3* construct, while one line of *ANAC055::WRKY3* had reduced lesion size at 48 and 72 hours post infection. However, for full confidence in these results, the infections should be repeated in all cases for T3 plants. Interestingly, *ANAC055::TGA3* rewirings seemed to show an increase in lesion size. *tga3* plants were clearly more susceptible to infection in Windram *et al.* (2012) - however this does not necessarily mean that upregulating it would decrease susceptibility to infection. In the *At-Botrytis* network, *WRKY53*, a negative regulator of defence, is directly regulated by TGA3; there may be other TFs and genes downstream of *TGA3* which have a negative impact on infection. This justifies the need for accurate and systematic models to accurately predict the effect of node perturbations on the network as a whole. Transcriptome measurements at a single, or several time points, would also assist in understanding the changes brought on by rewiring.

Interestingly, Isalan *et al.* (2008) did not find that *E. coli* transformants with a high copy number plasmid containing promoter::GFP fusions influenced the transcriptome of the cell. This is unexpected, as the unusually high levels of promoter could titre out transcription factors that normally bind these regions. It also means that modifying the network creates fewer unwanted side-effects. A general observation of rewiring is that similarly to Isalan *et al.* (2008) and Windram *et al.* (2017), the level of expression of a rewired gene depends not only on the regulatory region it was fused to, but also the coding region of the gene itself. This is probably due to downstream positive and negative motifs. These TFs do not work in isolation, unlike GFP, and potentially interact with the promoter regions of downstream TFs, which then regulate the original TF, resulting in feedback loops. With Golden Gate cloning, it is possible to piece together several rewiring constructs from the existing plasmids. Therefore several network rewirings that result in decreased lesion size can be combined into a single line. This may necessary to counter the robustness of biological network to perturbations (Isalan *et al.*, 2008).

As the disparity in gene expression between protoplasts and leaves is great, another method for high throughput screening of constructs would be beneficial. The transient rewiring of single Arabidopsis leaves using *A. tumefaciens* as introduced by Tsuda *et al.* (2012), followed by *B. cinerea* infection, is one such alternative. *A. tumefaciens* infiltration of Arabidopsis leaves is a challenge, unlike the ease which comes with transiently transforming *N. benthamiana*. By using Arabidopsis expressing the *P. syringae* effector AvrPto prior to transformation, Tsuda *et al.* (2012) achieved a large increase in transformation efficiency of inoculated leaves. They demonstrated this by transforming Col-0 and GVG-AvrPto (Hauck *et al.*, 2003) Arabidopsis with plasmids containing *35s::GUS* and performing a GUS assay or a HRB1-cyan fluorescent fusion and monitoring the protein's subcellular localisation. AvrPto appears to function by suppressing papillae-associated cell wall defence (Hauck *et al.*, 2003), possibly by blocking the kinase activity of the receptor-kinase FLS2 (Xiang *et al.*, 2008). The ability of AvrPto to mimic the modulation of host gene expression, however, may affect the

outcome of *B. cinerea* infection through antagonistic interactions with the immune signalling response and TFs. Transient expression using GVG-AvrPto may function better in rewiring TFs for enhancing defence against *P. syringae*, as AvrPto induces similar gene expression to the type 3 secretion system.

Overall, while protoplasts provide great advantages in terms of ease of use and high throughput screening, they would require further characterisation and great care when using as a substitute for a whole plant system.

5.6 Materials and Methods

5.6.1 RNAseq analysis

Five control and five chitin-treated protoplast samples were submitted for RNA sequencing. Data was obtained from the Oxford Genomics Centre in fastq format, with read 1 and read 2 of the paired end reads for each sample. All analysis was carried out on the York Advanced Research Computing Cluster. Prior to RNAseq analysis, quality control of the results was carried out using FastQC (Andrews *et al.*, 2010) to determine whether discarding of some samples or trimming of reads would be required:

```
$ find reads/ -name '*.fastq.gz' -exec fastqc {} -o ./fastqc \;
```

The quality score across all positions in reads from all 10 samples were good, therefore the trimming step was skipped.

kallisto was used to construct an index of the Arabidopsis Araport 11 reference transcriptome (Cheng *et al.*, 2017):

```
$ kallisto index -i transcripts.idx --make-unique araport11.fasta
```

The read data was then aligned to this index using pseudoaligment and the transcript abundances quantified:

```
$ kallisto quant -i transcripts.idx -b 100 sample1_read1.fastq.gz
sample1_read2.fastq.gz
```

Transcript abundance data was imported into R (Team *et al.*, 2013) using tximport (Soneson *et al.*, 2015). Transcripts with fewer than 100 counts across all 10 samples combined were eliminated. If gene counts were needed for analysis, they were generated from abundances so that they were scaled using the average transcript length and to library size, and summarised on a gene- rather than transcript-level using tximport.

For determining differential gene expression, the edgeR (Robinson *et al.*, 2010) function DGEList was first used to create a features table for the RNAseq data. Normalisation factors for library size are then calculated using the weighted trimmed mean of M-values (Robinson and Oshlack, 2010) and the edgeR function calcNormFactors. limma-voom (Law *et al.*, 2014) was finally used to transform count data to log2-counts

per million, fit a linear model for each gene and calculate empirical Bayes statistics for differential expression between the control and treated samples.

R code for the steps transcript abundance manipulation and quantification of differential gene expression is given in Appendix D.

5.6.2 GO term analysis

GO enrichment analysis of biological process terms was performed with the PANTHER Overrepresentation Test (Mi *et al.*, 2013). Only biological process terms with a p-value < 0.05 in a Fisher's exact test with the false discovery rate adjusted for multiple test correction were retained.

5.6.3 Arabidopsis lines

Arabidopsis Columbia lines were obtained through the Nottingham Arabidopsis Stock Centre (NASC) or from the original authors as outlined in Table 5.7. Arabidopsis ecotype Columbia-0 (Col-0) is referred to as WT Arabidopsis in this thesis.

Background	Mutation type	Mutated Gene/Insert	Source
Col-0	Knock-out	*bos1*	Mengiste *et al.*, 2003
Col-0	Insertion	*35S::GFP*	Earley *et al.*, 2006
Col-4	Knock-out	*at-erf1*	PRESTA project

Table 5.7: Sources for Arabidopsis seeds used in this thesis. *The Arabidopsis line with *GFP* under the control of the *35S* promoter was made by the Beynon group at Warwick University using a pEarley Gate plasmid from Earley *et al.* (2006).

5.6.4 Plant growth conditions

Arabidopsis thaliana seeds were stratified in soil for 24 hours in the dark at 4°C before being grown in a 12:12 hour light:dark cycle in a plant growth chamber at 22°C, 70% humidity and light intensity of 100 μmol photonsm^{-2}s^{-1}. Arabidopsis soil mix used was Levington F2+s compost (seed and modular compost with sand) and the plants were grown in P24 Desch-Plant-pak pots in propagator trays. The P24 pots were covered with a plastic cover to ensure a humid environment for the seedlings for the first week. Seedlings were transplanted so that only one seedling was grown per pot.

5.6.5 Isolation of Arabidopsis leaf protoplasts

Arabidopsis of Columbia 0 ecotype (Col-0) were grown in a Percival growth chamber for 3 to 4 weeks in the conditions described above. 3 of the largest leaves were harvested from each plant and cut into 1 mm-thick strips.The leaf material was transferred to 6 ml of filter-sterilised enzyme solution (20 mM MES at pH 5.7, 0.4 M mannitol, 20 mM KCl, 1.5% weight/volume Cellulase R-10 from Melford and 0.4% weight/volume Macerozyme R-10 from Melford) in a Petri dish. The enzyme solution was prepared fresh and stored in 6 ml aliquots at -20°C. The solution was vaccuum infiltrated for 30 seconds and the desiccator was placed in the dark for a further 15 minutes. The leaves were then left to digest in the dark for 2 hours. All subsequent handling of protoplasts was done with cut pipette tips. The solution was diluted with an equal volume of cold W5 solution (2 mM MES pH 5.7, 154 mM NaCl, 125 mM $CaCl_2$ and 5 mM KCl) and filtered using a reusable mesh into a new 50 ml tube to remove leaf strips. The resulting protoplast solution was centrifuged at 100 g in a desktop Fresco 21 centrifuge for 1 minute. The supernatant was discarded and the protoplasts resuspended in 5 ml of W1 solution (4 mM MES, pH 5.7, 0.5 M mannitol and 20 mM KCl).

Protoplast viability was verified under a light microscope and the number of live protoplasts (with an intact cell membrane) was counted using a Fuchs-Rosenthal counting chamber. A more accurate accounting of live protoplasts was achieved by staining for 5 minutes with a Fluorescein Diacetate stain (1 μl in 30 μl total volume of protoplast solution) and visualising with a Zeiss Axiovert 200 microscope.

5.6.6 Protoplast transformation

Protoplasts were centrifuged at 100 g for 1 minute and resuspended at a concentration of $3 * 10^5$ protoplasts/ml in MMG solution (4 mM MES pH 5.7, 0.4 M mannitol and 15 mM MgCl2). For triplicates of single transformations, 3 μl or 6 μl of 1 μg/μl plasmid construct solution were added to a 2 ml Eppendorf tube. 90 μl of protoplast solution was transferred to each tube and an equal volume of fresh PEG-calcium transfection solution (40% wt/vol PEG4000, 0.2 M mannitol and 100 mM $CaCl_2$) added. The solution was shaken at 1000 rpm for 1 minute, followed by incubation at room temperature for 15 minutes. 510 μl of W5 solution (2 mM MES pH 5.7, 154 mM NaCl, 125 mM $CaCl_2$ and 5 mM KCl) was added and mixed by shaking at 1000 rpm for 1 minute. The solution was centrifuged for 2 minutes at 100 g with decreased acceleration and deceleration. The supernatant was removed and replaced with 450 μl of W1 solution and the protoplasts resuspended by shaking for 1 minute at 1000 rpm. 150 μl of protoplast solution was added to each of 3 wells on a white 96-well plate which was covered with breathable film and left to incubate overnight in the dark.

5.6.7 Chitin induction of protoplasts

A chitin stock solution of 1 g/ml was prepared by autoclaving 5 g of chitin from shrimp shells (Sigma) in 5 ml of water. Chitin solutions of 500, 100 and 50 mg/ml

concentration were prepared by serial dilution in protoplast buffer W1. 100 μl of the dilute chitin solution was added to 1 ml of protoplast suspension, giving a final chitin concentration of 5, 10 or 50 mg/ml. Mock inoculum consisted of buffer W1.

Protoplasts were induced with chitin or W1 buffer as the control for 15, 30, 45 or 60 minutes. After induction with chitin or mock inoculum, the protoplasts were snap frozen with liquid nitrogen and stored at -80°C until needed for RNA extraction.

5.6.8 RNA extraction

5.6.8.1 RNA extraction from Arabidopsis protoplasts

Total RNA isolation from protoplasts was performed using the QIAGEN RNeasy Plant Mini Kit according to the manufacturer's protocol, but with the disruption of plant material step omitted. RNA was eluted in 35 μl of DEPC-treated water. The concentration and purity of RNA was determined using a NanoDrop spectrophotometer and the integrity of the RNA was determined by running 1 μl of RNA on a 1% agarose gel or using the Agilent 2100 BioAnalyser (for samples meant for RNAseq).

If the RNA samples had the required RIN> 7.5, they were sent for RNAseq. For samples to be used for qPCR, if the NanoDrop peaks showed no trace of contamination and the gel showed no degradation, the samples were treated with DNase I from Sigma according to the manufacturer's protocol before proceeding with cDNA synthesis.

5.6.8.2 RNA extraction from Arabidopsis leaves

Total RNA isolation from leaves was performed using the QIAGEN RNeasy Plant Mini Kit according to the manufacturer's protocol. Two 6mm glass beads were added to Eppendorf tubes containing leaf samples, which were then homogenised using a Mixer Mill (MM 400) for 30 seconds. The grinding process was repeated, with the tubes on the outside of the adapter racks moved to the inside, and vice versa. Samples were flash frozen again to prevent thawing before the subsequent steps were carried out. RLT Buffer was then added to the samples and the extraction continued following the protocol. RNA was eluted in 35 μl of DEPC-treated water. Samples were then treated with DNase I from Sigma according to the manufacturer's protocol before being used for cDNA synthesis.

5.6.9 cDNA synthesis and qPCR

cDNA synthesis was carried out using Invitrogen SuperScript IV Reverse Transcriptase following the manufacturer's protocol. The cDNA was then used as a template for real-time quantitative PCR. For protoplast RNA, the cDNA was used without dilution in qPCR; for RNA extracted from leaves, the cDNA was diluted 10-fold before being used in a qPCR experiment.

qPCR reactions were carried out using PrimerDesign PrecisionPLUS qPCR Master Mix with ROX at a reduced level premixed with SYBRgreen in a QuantStudio3 Real Time PCR System or Bio Rad CFX384 Real Tme PCR Detection System. 10 μl of

Table 5.8: Thermocycling conditions used in qPCR reactions.

Step	Temperature, °C	Time, min:s	Notes
1	95	2:00	
2	95	0:10	40 cycles
	60	1:00	
3	95	0:15	Melt curve
	60	1:00	
	95	0:01	

Master Mix was added to 1 μl of cDNA, 1 μl of primer mix and 8 μl of sterile water. Samples were then pipetted into an Applied Biosystems MicroAmp Fast 96-well or a Thermo Scientific white 384-well reaction plate. The thermal profile of the qPCR reaction is described in Table 5.8. The melt curve conditions used the instrument default settings. Only primer pairs that had a single peak in the melt curve were used for the final experiments to ensure amplicon specificity.

The primer sequences for amplifying genes are listed in Table 5.9. If the gene contained introns, primers were designed using NCBI Primer-BLAST (Ye *et al.*, 2012) and made to span exon-exon junctions. The length of the final product was made to span from 50 to 200 base pairs.

qPCR results were visualised with the QuantStudio Design and Analysis cloud Software from the ThermoFisher website. The Rn threshold was manually selected for each experiment to obtain cycle threshold values above background fluorescence and in the exponential phase of all the samples. The normalised values of all genes for each condition were calculated relative to the housekeeping gene(s) using the formula:

$$\Delta\mathrm{Ct}_{\text{gene of interest}} = 2^{\mathrm{Ct}_{\text{housekeeping gene}} - \mathrm{Ct}_{\text{gene of interest}}} \tag{5.1}$$

The expression fold change of a gene after treatment, the $\Delta\Delta$Ct, was calculated via:

$$\Delta\Delta\mathrm{Ct}_{\text{gene of interest}} = \Delta\mathrm{Ct}^{\text{treatment}}_{\text{gene of interest}} / \Delta\mathrm{Ct}^{\text{control}}_{\text{gene of interest}} \tag{5.2}$$

5.6.10 RNA-Seq library preparation

The library preparation was performed by Dr Sally James at the University of York Technology Facility. Poly(A) purification of samples with the NEBNext Poly(A) mRNA Magnetic Isolation Module was followed by library preparation with the NEBNext UltraII directional RNA sequencing kit. The 10 samples were indexed with the first 10 adapters from the TruSeq Stranded mRNA LT Adapter AR series, and pooled at an equimolar ratio. The resulting pool had a concentration of 17 ng/μl and an average library size of 370 base pairs.

Table 5.9: Primer sequences for the amplification of genes during qPCR.

AGI	Gene name	Primer sequence F	Primer sequence R	Amplicon size, bp	Span exon-exon junction?
AT5G05610	*AL1*	TCGCAATGAGAGGAAACGTC	TCCATTAGTTTCGGCGTGCT	180	Yes
AT3G56530	*ANAC064*	TGGAGGATATTGGAGCGCAA	GCATGATCCACCTTCTCGTT	186	Yes
AT5G13180	*ANAC083*	GCAAACCACCTCATGGCTCA	ACCCATAGAACTCGGAGGAGA	80	Yes
AT1G59750	*ARF1*	GCTAAAGGTTCAATGGGACGA	GGGATTTCCACACACCATCA	191	Yes
AT4G30080	*ARF16*	CTTCTACAGGTGGCGTGGGA	GCGAAAACGAAGTAAGCGGG	115	Yes
AT1G56060	*ATHCYSTM3*	CAGCATGAAATGAAGCAGACA	CGTAAAAATCCACCACTGCC	146	Yes
AT5G57150	*bHLH*	TCTCATTCCCCATTCGCGTT	AAAGGCCAGCTTCTGTCGTA	199	Yes
AT2G22300	*CAMTA3*	AAGCAAGACGATTCAGCCCA	ACAGACCCACTTGATGGTGT	163	Yes
AT3G23240	*ERF1*	CTAATCGAGCAGTCCACGCA	GTCCCGAGCCAAACCCTAAT	182	No
-	*GFP*	ATCATGGCCGACAAGCAGAA	TCTCGTTGGGGTCTTTGCTC	187	No
AT5G15770	*GNA1*	GGAAAAATCGCTGCTACGGG	GCCCAGCTTTACCGCAATTC	70	No
AT4G31270	*IDAP2*	CCGGCGATCTGTTTAAGCC	TCAACCGCAGCGATCTCATT	130	Yes
AT1G72520	*LOX4*	ACTTCCCCTGCAACTCTTGG	GCTCGGCAAGTAAGGCTGA	86	Yes
AT5G44420	*PDF1.2*	AAGTTTGCTTCCATCATCACCC	ATTGCCGGTGCGTCGAAAG	71	Yes
AT1G76580	*SBP*	TGCACAAGGTAGGAACGAGG	GGATTCATCGGGGAGGGTTG	180	Yes
AT1G22070	*TGA3*	GAGTTTTGTGAACCAGGCGG	AGAACTAAGAGCACGCAGCC	136	Yes
AT4G14960	*TUA6*	TGGCAAGATGAGCACAAAAG	AGACCTCGGGGAGCTATG	129	No
AT3G62250	*UBQ5*	AAGAAGACTTACACCAAGCCGAAG	ACAGCGAGCTTAACCTTCTTATGC	62	No
AT2G03340	*WRKY3*	AGCAGAGTAGACCAACGGGA	ACCAAATGTTCCCTGAAGAGG	137	Yes
AT2G38470	*WRKY33*	AGCAAAGAGATGGAAAGGGGA	TGCACTACGATTCTCGGCTC	87	Yes
AT2G25000	*WRKY60*	ATTGGCTCCTGTGAACCTAAAT	CTCTTTTCGGAAGTTTACCGGAG	88	Yes
AT3G56400	*WRKY70*	TGCCAAATTCCCAAGAAGTTAC	TTGGGAGTTTCTGCGTTGGT	156	Yes
AT5G46350	*WRKY8*	AACAGTCCTTATCCGAGGAGTT	ATCAGATGCGGTGGTTCCAG	180	No

Table 5.10: Plasmids used for Golden Gate reactions.

Plasmid	Plasmid number	5' Overhang	3' Overhang
L0 promoter acceptor	pICH41295	GGAG	AATG
L0 CDS acceptor	pICH41308	AATG	GCTT
35S Terminator	pICH41414	GCTT	CGCT
L1 acceptor	pICSL86955OD	GGAG	CGCT

Sequencing was carried out in the Oxford Genomics Centre at the Wellcome Centre for Human Genetics. The library was sequenced on a HiSeq 4000 Illumina machine which returned 75 paired end reads in FASTQ format.

5.6.11 Golden Gate cloning

Plasmids for Golden Gate cloning (Table 5.10) were kindly given by Mr. Mark Youles from The Sainsbury Laboratory. Level 0 acceptors (including the plasmid containing the terminator of translation) carried a spectinomycin resistance gene and the Level 1 acceptor carried a kanamycin resistance gene. The plasmids all contain a *lacZα* gene in the insertion site for blue-white screening. In addition, the level 1 acceptor also carried a *bar* gene for BASTA selection in Arabidopsis plants.

5.6.11.1 Part amplification

Promoter regions consisting of 2000 base pairs upstream of the ATG start site were amplified from genomic DNA (gDNA). Arabidopsis gDNA was extracted using the Extract-N-Amp Plant Tissue PCR Kit (Sigma) following the manufacturer's protocol. The primers for the promoter regions were designed to contain GGAG-AATG overhangs flanking the promoter sequence.

cDNA or gene coding regions (CDS) were amplified from Arabidopsis cDNA, with the exception of GFP, which was amplified from a pEarleyGate 201 plasmid containing GFP (Earley *et al.*, 2006). The primers for the coding regions were designed to include AATG-GCTT overhangs flanking the CDS. A number of parts were amplified in 2 fragments if they contained BsaI or BpiI recognition sites in a process known as domestication. BsaI and BpiI recognition sites are: GGTCTC and GAAGAC, respectively. The primers for domestication were designed to include a silent mutation in the protein and remove the enzyme recognition site. The primers for amplifying DNA fragments are shown in Table 5.11.

Table 5.11: Primers for amplifying fragments for Golden Gate cloning.

Gene	Fragment	F/R	Primer
AL1	1	F	AAGGTCTCTAATGGCCGCTGAGTCTTCTAACCC
AL1	1	R	ATGGTCTCAAAGCGAGACTGCAGGAGATGAGGT
AtWHY2	1	F	ATGAAGACATAATGATGAAGCAAGCCCGCTCTTT
AtWHY2	1	R	ATGAAGACGGGTCCTCATCACTGCAAATTCAGCTTTTG
AtWHY2	2	F	GGGAAGACATGGACAGCTTTTAGTTTTGCTCTTCCA
AtWHY2	2	R	ATGAAGACATAAGCTCATTTATCCCACTCCAGCT
TGA3	1	F	ATGAAGACATAATGGAGATGATGAGCTCTTCTTC
TGA3	1	R	ATGAAGACATGTCCTCGTCCTGATCATTATTACT
TGA3	2	F	GGGAAGACATGGACCGGATCAATGATAAGATGAAACG
TGA3	2	R	ATGAAGACATAAGCTCAAGTGTGTTCTCGTGGAC
WRKY3	1	F	ATGAAGACATAATGGCGGAGAAGGAAGAAAAA
WRKY3	1	R	ATGAAGACATAGATCTAGGTGCAGAAGAAAATGA
WRKY3	2	F	ATGAAGACATATCTCAGATTCGAGCCTCGGTT
WRKY3	2	R	AGGAAGACGGAAGCTCAAGTGATTTGCTCTTCTT
WRKY60	1	F	ATGAAGACATAATGGACTATGATCCCAACACCA
WRKY60	1	R	ATGAAGACATAAGCTCATGTTCTTGAATGCTCTA
WRKY70	1	F	ATGAAGACGGAATGGATACTAATAAAGCAAAAAAGCT
WRKY70	1	R	ATGAAGACATGTCCTCTTCCTTCATTGAGGTAGA
ANAC055	1	F	ATGAAGACATGGAGGCATTCGAGTTCGAATTTCA
ANAC055	1	R	ATGAAGACATGTCATCCCAAGTATTAGGGAGTTA
ANAC055	2	F	ATGAAGACGGTGACCATTGACAATACAAAAAAAA
ANAC055	2	R	ATGAAGACATCATTTTTTCCTATCACCTCTTT
SAP12	1	F	ATGAAGACATGGAGCATTCAACAGTTTGAATCAA
SAP12	1	R	GCGAAGACGGCATTAAGCTTCTCTTCTTCTTCTT
ZAT11	1	F	ATGAAGACATGGAGAAGAACTTAATGTCACAATGG
ZAT11	1	R	ATGAAGACATAGATCAAACAAAAGCCTTTTTGAT
ZAT11	2	F	ATGAAGACATATCTCTAGTGCATGTGTTTGACCA
ZAT11	2	R	AGGAAGACGGCATTACTTCTTTGAGAATTAAGATCTGAG
GFP	1	F	ATGAAGACTAAATGGTGAGCAAGGGCGAGGAGCT
GFP	1	R	ACGAAGACGGAAGCTTACTTGTACAGCTCGTCCA

Table 5.12: Thermocycling conditions for Golden Gate cloning reactions.

Step	Temperature, °C	Time, min:s	Notes
1	37	0:20	
2	37	3:00	26 cycles
	16	4:00	
	50	5:00	
3	80	5:00	

5.6.11.2 Gel electrophoresis and DNA purification

DNA was visualised on a 1% agarose gel stained with 1 μl ethidium bromide per 100 ml of gel. The gel consisted of AGTC Bioproducts agarose dissolved in 0.5X TBE buffer (1 mM EDTA, 45 mM Tris-Boric acid pH 8.0. An NEB 100 bp or 1 kb Plus DNA ladder was used, depending on the size of expected products. Bands of the expected size were cut out from the gel with a sterile scalpel and purified using the QIAquick Gel Extraction Kit (QIAGEN) according to the manufacturer's protocol.

5.6.11.3 Golden Gate cloning

A Golden Gate Level 0 reaction was carried out to insert a CDS or promoter region into their respective plasmid. 100 ng of acceptor plasmid was combined with the insert(s) at a molar ratio of 2:1 insert:acceptor, 1.5 μl T4 DNA Ligase Buffer (NEB), 1.5 μl of bovine serum albumin diluted 10x, 0.5 μl of T4 DNA Ligase (400 U/μl, NEB), and 0.5 μl BpiI (ThermoFisher) and made up to 20 μl with sterile water. These were then amplified according to the thermocycling conditions in Table 5.12.

Golden Gate Level 1 reactions were used to excise promoter and CDS regions from their acceptor plasmids and assemble them in the Level 1 acceptor. 100 ng of the acceptor plasmid was combined with the plasmids containing the insert at a molar ratio of 2:1 insert:acceptor, 1.5 μl T4 DNA Ligase Buffer (NEB), 1.5 μl of bovine serum albumin diluted 10x, 0.5 μl of T4 DNA Ligase (400 U/μl, NEB), and 0.5 μl BsaI (ThermoFisher) and made up to 20 μl with sterile water. These were also amplified according to the thermocycling protocol in Table 5.12.

5.6.11.4 *E. coli* transformation

The products of Golden Gate Level 0 or Level 1 reactions were transformed into High Efficiency DH10B Competent *E. coli* (NEB) and Library Efficiency DH5 α competent cells (ThermoFisher), respectively, in order to select for the plasmids with the correct insertions. 5 μl of the Golden Gate reaction was mixed with 30 μl of competent cells, and the transformation was carried out according to the manufacturer's protocol.

Plates for *E. coli* blue-white screening were prepared by autoclaving pre-mixed Luria-Bertani (LB) agar powder (Formedium) in sterile water. The desired antibiotic was added when the media cooled down to 60°C and ∼10 ml were poured into 60 mm x 15 mm sterile plates. Kanamycin and spectinomycin were both diluted to a working concentration of 50 μg/ml. The LB agar was left to solidify. 40 μl of 100mM IPTG (Promega) and 120 μl of X-Gal (Thermo Fisher Scientific, 20 mg/ml) were added to the surface of each plate and spread over the entire surface with a sterile plastic spreader. The plates were again left to dry before streaking out 10-100 μl of *E. coli* solution and incubating overnight at 37°C.

Colonies that were white and grew on spectinomycin had successfully incorporated the CDS or promoter into their respective plasmid acceptors in the Level 0 reaction. Colonies that were white and grew on kanamycin had successfully incorporated the promoter-gene construct into the Level 1 plasmid acceptor.

5.6.11.5 Plasmid purification

If small amounts of plasmid needed to be purified, the QIAprep Spin Miniprep kit (QIAGEN) was used. Single *E. coli* colonies were transferred from an LB plate into 5 ml of liquid LB with the correct antibiotic (50 μg/ml of kanamycin or spectinomycin). These were grown for 16 hours in a 37°C shaking incubator at 280 rpm. The bacterial cells were harvested by centrifuging at 5000 rpm in a table-top centrifuge at room temperature for 10 minutes. The manufacturer's protocol was followed for the rest of the process, and the plasmid DNA was eluted in 50 μl Buffer EB.

If larger concentrations of plasmid were required, the QIAGEN Midiprep kit was used. Single *E. coli* colonies were grown in 5 ml of liquid LB with antibiotic (50 μg/ml for 16 hours in a 37°C shaking incubator). 1 ml of this liquid culture was used to inoculate 50 ml of fresh LB medium in a 250 ml sterile flask and grown overnight in a 37°C shaking incubator. The bacteria were harvested by centrifuging at 5000 rpm in a table-top centrifuge at 4°C for 15 minutes. The manufacturer's protocol was followed for the rest of the process, and the plasmid DNA was eluted in 500 μl of 10 mM Tris-HCl, pH 8.5 buffer.

5.6.12 *Agrobacterium tumefaciens*-mediated transformation of Arabidopsis

5.6.12.1 *Agrobacterium tumefaciens* transformation

100 μl of *Agrobacterium tumefaciens* GV3101 was thawed on ice. 1 μg of plasmid was added to the cells and mixed. The mixture was incubated on ice for 5 minutes and heat shocked by freezing in liquid nitrogen for 5 minutes. The tubes were then transferred to a 37°C water bath for 5 minutes. The cells were incubated on ice for 2 minutes, after which 900 μl of YEB liquid medium (5 g/l beef extract, 1 g/l yeast extract, 5 g/l peptone, 5 g/l sucrose, and 2 mM $MgSO_4$ made up in sterile water and autoclaved) was added. The cells were incubated for 2-3 hours in a 28°shaking incubator. Following

this, the cells were centrifuged at 3000 rpm in a desktop microfuge for 30 seconds. 900 μl of the medium was removed and the cells were resuspended in the remaining medium. The cells were streaked onto a YEB agar plate containing 100 μg/ml rifampicin, 50 μg/ml gentamycin and 50 μg/ml kanamycin (the latter for the plasmid). YEB agar plates were prepared using the same recipe as for YEB media, with the addition of 15 g/l of agar. The antibiotics were added to the autoclaved media once it had cooled down to 60°C and ~10 ml of media were poured into 60 mm x 15 mm sterile plates. The plates were incubated for 24-48 hours at 28°C.

5.6.12.2 Arabidopsis transformation by floral dipping

Single *Agrobacterium* colonies were picked from YEB agar plates using a sterile loop and grown in 10 ml of YEB liquid medium containing rifampicin, gentamycin and kanamycin (in the same concentrations as Section 5.6.12.1) overnight in a 28°C shaking incubator. 500 ml of YEB media with antibiotics was prepared by autoclaving in autoclaved 2 l flasks. 5 ml of the 10 ml culture was added to the 500 ml YEB and incubated overnight in a 28°C shaking incubator. The bacterial cells were harvested by centrifuging for 10 minutes in 250 ml Nalgene centrifuge bottles in a High Speed Sorvall Evolution 2 with a F10 6x500 rotor at 4°C.The supernatant was decanted and the cells were resuspended in 1 l of 5% sucrose solution with 200 μl of Silwet (PlantMedia). The solution with *Agrobacterium* was poured into 2 l plastic beakers. Arabidopsis Col-0 plants was grown until just flowering with many young unopened flower buds (6-7 weeks). 4 plants were held together, and the buds were immersed in the bacterial culture and swirled gently for 30 seconds. The plants were set on their sides in propagator trays which were placed inside autoclave bags and sealed. After 1 day, the plants were stood upright and left to grow in a growth chamber as normal (Section 5.6.4).

5.6.12.3 BASTA selection of transformed Arabidopsis

Seeds gathered from transformed Arabidopsis plants were stratified for 24 hours in the dark at 4°C on soil soaked in BASTA herbicide (Bayer, 1 ml of the herbicide to 1 litre of water). As many seeds as possible were sown in a P24 Desch-Plant-pak pot tray without overcrowding, as very few of the primary transformants will be transgenic. These were left in a growth chamber for 10 days without watering. Col-0 untransformed seeds and a pEarleyGate 201 plasmid containing GFP (Earley *et al.*, 2006) were used as the negative and positive control, respectively. Seedlings that grew on BASTA-treated soil were transgenic and moved onto normal F2+s compost after 10 days, and grown as normal.

5.6.13 *Botrytis cinerea* culturing and infection assays

Botrytis cinerea strain pepper (Denby *et al.*, 2004) was cultured on tinned apricot halves in petri dishes in a cabinet in constant darkness at 25°C. Subculturing was carried out every 3 weeks. Spores were harvested in 3ml of sterile water from the apricot halves

and filtered through glass wool to remove hyphae. The number of spores per ml was quantified using a haemocytometer, and diluted to a concentration of $2x10^5$ spores/ml. An equal amount of grape juice was added to give a final concentration of 10^5 spores/ml.

For *B. cinerea* infection assays, leaves from five-week-old Arabidopsis plants were detached and placed on 0.8% agar in propagator trays. A single 10μl droplet of fungal inoculum was pipetted onto each leaf. Mock-treated leaves had grape juice diluted in an equal amount of sterile water as the inoculum. Trays were covered with a plastc lid and placed in a cabinet with the same conditions as those for growing plants, except the humidity was raised to 90%. Leaves were inoculated 8 hours after dawn. Photographs were taken of the leaves at 48, 60 or 62 and 72 hours post inoculation. These were analysed with Fiji (Schindelin *et al.*, 2012) using a 5cm scale measure placed in each tray to determine the area of each leaf lesion. The macro for calculating lesion area is give in Dataset G. A Student's two-tailed t-test was used to compare the lesion areas of control lines and genetically modified lines.

Leaf samples for gene expression analysis were inoculated with spores as described above, except that five 7μl droplets were applied evenly spaced over the leaf. The leaves were collected in tubes 28 hours post inoculation, flash frozen in liquid nitrogen and stored in a -80°C freezer until RNA extraction.

Chapter 6

Discussion

The interaction between plants and pathogens is a continual arms race where reprogramming of the plant transcriptome plays a large part in the outcome of disease. This interdisciplinary thesis focuses on deciphering these complex regulatory interactions that take place in Arabidopsis during infection with the necrotrophic fungal pathogen *Botrytis cinerea*. Secondly, manipulation of the *Arabidopsis thaliana* transcription factor (TF) network through *in silico* and *in vivo* rewiring is employed with the aim of enhancing the plant's defence response. Rewiring is a method much-used by organisms during their development and evolution. Its potential as a synthetic biology tool has been explored here. Rewiring reroutes signals and short circuits processes, resulting in a fitness advantage under certain circumstances. In order to achieve this goal, network inference and analysis have been used in conjunction with control engineering, Gaussian process dynamical systems modelling, and the transient and stable transformation of Arabidopsis plants. All rewiring was limited to promoters and genes that are already present in Arabidopsis. This has a number of practical advantages: it eschews the need for characterising orthogonal parts (parts foreign to the system, which cannot be recognised by it or interact with it), is guaranteed to have proper folding and modification of the protein, and may encounter fewer regulatory constraints. Rewiring, rather than over-expressing or knocking out native genes, is a more subtle way of affecting gene expression and is likely to have a smaller impact on the regular functioning of the plant. Chapters 2-4 describe approaches to formulating testable hypotheses and are applicable to other organisms, as long as the prerequisite data collection has taken place. Chapter 5 presents a methodology for experimental validation of these network-based hypotheses in Arabidopsis.

Rewiring techniques have also been implemented by others for various purposes. Isalan *et al.* (2008) studied the consequences of inserting a construct containing a promoter::gene rewiring in *E. coli*. All pairwise combinations of 27 promoters and 23 genes were attempted systematically, with 598 made and successfully used for the transformation. $\sim$ 95% of genetically modified *E. coli* were viable, emphasising the robustness of the microbe to regulatory changes. Out of two constructs whose transcriptome was measured with microarrays, rpoS-ompR had only 13 DEG, implying that some rewirings

do not propagate change widely. These small-scale rewiring events are a cornerstone of evolution (Borneman *et al.*, 2007; Gompel *et al.*, 2005; Prud'Homme *et al.*, 2006), so it is unsurprising that they are well-tolerated. Due to their prominence in the evolution process, Isalan *et al.* (2008) were interested in the effect these rewirings had on survival in regular and extreme conditions. Some rewired clones were repeatedly selected for during serial passaging in liquid culture, survival during extended periods at 37°C and heat shock survival, particularly rpoS-ompR or ones containing flhD promoters. flhD controls the expression of flagellar genes, so perhaps the rewiring affects the expression of these genes, freeing up resource to combat stress.

A follow-on study from Isalan *et al.* (2008) took a more in-depth look at the gene expression profiles of 85 *E. coli* rewirings with microarrays (Baumstark *et al.*, 2015). The number of differentially expressed genes from each rewiring ranged from 0 to 3 orders of magnitude. When well-connected TFs are transformed into cells under a different promoter, the transcriptome is more likely to have bigger perturbations. Degree of a TF node accounts for 37% of the variability observed in the mean transcription perturbation. An additional ∼30% of the variability is explained by the promoter. The best way to create constructs with low perturbation is to link together TFs with a low degree with weak promoters. However, most promoter::GFP constructs show that, on their own, promoters are not able to alter transcriptomes, at least not by depleting levels of TFs. Genes also appear to have a large say in their own expression levels – when the rewirings were clustered by gene expression levels, clusters originating from the same rewired gene were very prominent. This was also observed in Section 5.4.4.1 of the experimental results chapter, where different genes with the same promoter had different expression levels, but the expression level of a gene with different promoters stayed fairly constant. The gene itself appears to have a lot of control over its own expression, possibly through internal regulatory regions.

For both Isalan *et al.* (2008) and Baumstark *et al.* (2015), a concluding question is how to rewire the system to achieve a certain target, whether it be optimised gene expression, maximised growth, or other metabolic changes. This work has attempted to answer this question and established the use of ODEs and GPDS simulations for optimising gene expression through rewiring.

6.1 Constructing gene regulatory networks specific to Arabidopsis stresses

Essential data for this work-flow is a high-resolution time series gene expression assay for genes that are to be included in the network. In order to increase the accuracy of the gene regulatory network (GRN), multiple different network inference algorithms should be used, and the consensus created in conjuction with transcriptomic data from knock-outs (KO) and over-expressors (OE) (preferably) and TF-DNA binding information as priors. As shown in Tables 2.3 and 2.4, network inference is more accurate on smaller networks. However, by only including a limited number of regulators and targets, the context of the subnetwork within the larger network will be missed. More

powerful and accurate algorithms are needed for inferring large networks accurately, and since the start of this project, several new methods for inferring networks have been published (such as Barman and Kwon (2017), Aibar *et al.* (2017), Marti-Marimon *et al.* (2018)). A range of "accessory" methods have also recently become available. For instance, ExRANGES integrates with other inference algorithms, such as GENIE3 and Inferelator, and increases their accuracy (Desai *et al.*, 2017). TF2Network uses Arabidopsis position weight matrices from various databases to identify regulators for functionally similar or co-expressed genes by looking at their shared transcription factor binding sites (Kulkarni *et al.*, 2017). These new methods could increase the accuracy of the models used in this work, and should be explored by those interested in constructing networks from gene expression datasets.

Network characteristics that have been previously suggested to associate with the importance of the role played by transcription factors did not identify TFs with a role in defence in this study. Neither degree, centrality, clustering coefficient or absolute expression change of a TF were able to predict the likelihood of its playing a negative or positive role in defence (Figure 2.10). Interestingly, a node's indegree (but not the outdegree) does seem to be able to indicate the amplitude of its expression (Figure 2.11). While network characteristics did not help in identifying TFs that influence the response to *B. cinerea*, the common stress network made by combining 7 consensus stress networks proved useful in identifying TFs with a role in the response to biotic or abiotic stresses.

6.2 Ordinary differential equation modelling and rewiring of an Arabidopsis stress subnetwork

Starting with the 883-node *At-Botrytis* GRN, a smaller subnetwork was selected in Chapter 3 for detailed modelling using ODEs. Screening of knock-out and overexpression mutant lines revealed a TF that played a positive role in defence (*CHE*) and a TF that played a negative role in defence (*at-ERF1*) against *Botrytis* within that subnetwork. *CHE* is notably downregulated during infection, so a feedback control strategy was proposed to restore its levels *in silico*. This proportional integral controller is based on the work of Harris *et al.* (2015), with one major difference - we proposed that the controller be designed purely by rewiring the existing network rather than introducing exogenous elements. First, an idealised version of the controller, consisting of a coherent feedforward motif of type I with an added negative feedback loop, was demonstrated to restore the levels of *CHE* while maintaining its oscillatory expression (Figure 3.8). This network motif was then constructed *in silico* by adding regulatory links from MYB51 and ORA59 to *CHE* and from CHE to *MYB51*. The resulting perturbation mitigation was not as successful as anticipated - the expression of *CHE* did increase, but was not fully restored to pre-infection levels. To solve this problem, caused by unmodelled dynamics affecting *MYB51* and too high a degradation rate of *ORA59*, two additional links had to be introduced to the network. This had the desired effect of ensuring the levels of *CHE* remained constant despite perturbations. While this circuit was not implemented *in vivo*, that would be the next logical step, and an experimental outline is suggested in

the Discussion section in Chapter 3.

6.3 Modelling and rewiring of networks using Gaussian process dynamical systems

Rewiring of larger networks was demonstrated using Gaussian process dynamical systems (GPDS) in Chapter 4. This method was first applied to *in silico* DREAM networks with known gold standards, in order to evaluate its predictive ability (Figures 4.12 and 4.16). A larger subnetwork of the *At-Botrytis* GRN was then identified. This was formed by using RNA-seq and microarray expression data to select only the genes that had similar expression profiles in chitin-induced protoplasts and *Botrytis*-infected leaves. Rewiring of this network of 70 genes (the 70GRN) was to be simulated using GPDS. GPDS models were used to learn the underlying functions of the 70GRN and predict its gene expression levels upon systematic rewiring of each TF using the regulatory region of all other TFs. A simple optimisation function was employed to find the rewirings which achieved the desired gene expression levels in the 70GRN by maximising the levels of positive regulators of defence (Figure 4.22), however, the simulations for the rewired networks did not vary greatly from the simulations for the WT network.

As modelling techniques become more powerful, the aim is to combine the inference of networks and network design into a single process, by simply inputting the observed gene expression and the desired gene expression.

6.4 Rewiring Arabidopsis gene regulatory networks *in planta*

Finally, description of and the results from implementing rewiring *in planta* using constructs generated with Golden Gate cloning is given in Chapter 5. These had the coding regions (CDS) of one of a number of TFs which are positive regulators of defence, but which are downregulated during *B. cinerea* infection. The promoter regions regulating these CDS originated from genes that were strongly upregulated during infection. 6 of these constructs were transformed into protoplasts and shown to increase the levels of the rewired gene compared to a promoter::*GFP* control. The expression levels of defence marker genes in the protoplasts were quantified with qPCR to predict if the rewirings would have the desired effect in plants (Figures 5.15 and 5.16). Arabidopsis plants were stably transformed with 6 constructs through floral dipping and the resulting T1 generation used for an infection assay and gene expression measurement. The rewired plants did not show as marked an increase in the rewired gene as the protoplasts did. The RNAseq dataset generated of control and chitin-induced protoplasts highlights something that has been suspected by the plant community for a long time but not proven explicitly: the protoplast transcriptome should not be assumed to be very similar to the leaf transcriptome. While not as high as in protoplasts, the expression of the rewired *TGA3, WRKY3* or *WRKY60* plants did increase 2- to 20-fold on average. One rewiring in particular also showed a robust increase in resistance in terms of reduced lesion area:

SAP12::WRKY3 (Figure 5.18). Interestingly, the levels of *WRKY3* in this mutant were not as high as in the *ANAC055::WRKY3* plants, but the latter did not show any significant change in infection phenotype (Figure 5.17). In addition, *SAP12::WRKY3* had decreased levels of the defensin *PDF1.2*, and similar levels to wild-type of the defence gene *WRKY33*. Therefore the mechanism of action through which this rewiring conveys resistance remains unknown.

The promoter::GFP constructs themselves can help us understand how much of the variability in gene expression is controlled by the promoter in Arabidopsis. This was explored in *E. coli* (Baumstark *et al.*, 2015; Kosuri *et al.*, 2013). Kosuri *et al.* (2013) explored the influence of promoters and ribosome binding sites on translation and transcription by constructing a combinatorial library of 111 ribosome binding sites and 114 promoters. A similar approach can be used in Arabidopsis by trialling promoters of various lengths - 500, 100, 1500 and 2000 base pairs, and including or excluding the 5' untranslated regions.

6.5 Future work

Further work based on the presented results includes improving the network models and GPDS simulations by incorporating additional data, as well as undertaking further characterisation of the Arabidopsis lines generated.

Improving or testing the accuracy of the consensus *At-Botrytis* network can be done by ordering Nottingham Arabidopsis Stock Centre KO seed, infecting the adult plants with *B. cinerea* and following the transcriptomic response over time with qPCR, Nanostring or microarray. These experiments can identify genes downstream of the KO or OE TF, unless these are masked by redundancy. Using a combination of results from KO/OE plants and ChIP-sequencing could also help distinguish between the direct and indirect edges inferred by network inference algorithms, and establish which model is better at predicting these. The data can also determine whether there is a pattern associated with direct and indirect regulatory links which can be exploited to help network inference algorithms distinguish between them.

The 7 hubs found in the Arabidopsis common stress network which are not very well characterised merit further investigation for their role in the abiotic or biotic stress response. If any orthologues exist in crop species such as *Brassica napus* or *Solanum lycopersicum* of these hub genes, or hub genes from individual stress networks, they could be investigated for their role in defence against necrotrophs, given this prior knowledge about their role in Arabidopsis. An interesting line of enquiry which was touched on briefly here, and which could be expanded on in future, is the dynamic rewiring of transcription factor networks as a response to different stresses. As can be seen in the model results from Section 2.3.5, these networks are not static. However, it is hard to determine the extent of the rewiring in each circumstance as the topology of the network under control conditions is unknown. Guo *et al.* (2007) use a hidden temporal exponential random graph model to capture this evolving topology over time. This may allow the discovery of the network that all the initial observed data come from, before the

various stress perturbations affected it. Guo *et al.* (2007) used their algorithm to infer a *Drosophila melanogaster* muscle subnetwork at various stages of development from the embryonic, larval and pupal stage to adulthood. The rewiring of the network appeared to cease once the pupal stage was reached.

Implementation of the controller demonstrated in chapter 3 would be a considerable challenge, however it is likely to produce many insights into the requirements of constructing synthetic controllers *in planta*. Thus far, no published fully *in vivo* synthetic integral controllers exist, although a recent bioRχiv paper (Lillacci *et al.*, 2017) describes the construction of an integral feedback controller in *E. coli* based on a previously published antithetic controller design (Briat *et al.*, 2016a). Furthermore, the 9GRN is useful system in of itself for testing out new hypotheses, such as the extent of the change observed in the GRN under different stresses, and at steady state.

It would be prudent to model networks in protoplasts and leaves separately, given what was found in this thesis regarding the large differences between them. This would increase the accuracy for GPDS modelling and ensure the data used to verify and test the models is specific to the network model in question.

On the experimental side, an RNAseq dataset was generated of treated and untreated protoplasts which merits further examination to understand the inherent biases present when using protoplasts for screening (Schaumberg *et al.*, 2015). 18 promoter::gene rewiring constructs were generated which can be used for screening in a protoplast system with a luciferase reporter system or transiently transformed leaves as described in the Chapter 5 discussion. In addition, 14 stably transformed Arabidopsis lines were generated from these rewirings, out of which only 6 were tested for a *Botrytis* defence phenotype, and these only at the T1 generation. All lines merit characterisation in terms of infection assays and gene expression measurements at the T3 stage. The rewiring constructs generated are only the start; Golden Gate allows the synthesis of many promoter::gene pairs which can be tested for their effect on stress or other plant responses. For instance, the 18 constructs were created in in less than 2 months; and this process can be highly parallelised. The transcriptomic results from network rewiring can then be fed back to the network models to improve their accuracy, and this may enable modelling of the network to predict further rewirings. More cycles of design-build-test would increase the accuracy of predictions of the models.

While the protoplast system was not as faithful to the original leaf infection as was hoped, there is another possible way of quickly screening for rewirings that affect defence. Hauck *et al.* (2003) published a method for transient transformation of Arabidopsis leaves using *A. tumefaciens*, which could then be infected with *B. cinerea*. However, testing out the effect of rewiring on plant phenotype need not be limited to *B. cinerea* infection. The response to other pathogens such as *P. syringae* or *H. arabidopsidis*, or abiotic stresses such as salt, cold and osmotic stress can be quantified. If one particular rewiring was found to result in an overwhelmingly positive response to *B. cinerea* infection, it would be beneficial to test the response of this plant to other stresses, to ensure that this rewiring did not hugely impact these responses. While leaf protoplasts induced with chitin have limited use as an assay for the leaf response to the necrotroph, they may

form a better system for screening candidates for *P. syringae* response (by using flg22 as the inducer) or abiotic stresses.

The ultimate aim of this study is to rationally engineer crop plants to be more resistant to biotic and abiotic stresses such as *B. cinerea* infection. Any good candidates identified in Arabidopsis can be used to screen for homologs in tomato (Consortium *et al.*, 2012) or rice (International, 2005). International (2005) found that 90% of Arabidopsis proteins have a corresponding homolog in rice. Arabidopsis, which is quick and simple to grow, is a good candidate for proof-of concept studies such as these which can then be extended to engineer crop plants. The material and methods generated in this study provide a means to study rewirings and expand our knowledge of the complex and dynamic regulatory links that control the Arabidopsis response to external factors; they are also synthetic biology tools for engineering the desired response to a particular situation.

Chapter 7

Appendices

7.1 Appendix A - Network inference algorithms

7.1.1 CSI

CSI looks for dependence between two random variables: the levels of expression of a gene and the levels of expression of parent transcription factors that regulate this gene. It relates the inputs $\mathbf{x}_i$, which are the transcript levels of the parents at a previous time point, with the outputs y_i, which are the current transcript levels of the gene under consideration, using some function $f(\mathbf{x}_i)$:

$$y_i = f(\mathbf{x}_i) + \mathcal{N}(0, \sigma_n^2), \tag{7.1}$$

where $N(0, \sigma_n^2)$ represents some Gaussian noise, y_i are the levels of gene i and $\mathbf{x}_i$ are the levels of the parents of gene i.

By assigning the function $f(\mathbf{x}_i)$ a Gaussian process prior, the non-linear relationship between the inputs and outputs can be elucidated (Rasmussen and Williams, 2006). The inputs $\mathbf{x}_i$ and outputs y_i are expanded to form a matrix of inputs of parents at the previous time point $t-1$ which correspond to a vector of outputs at the current time point t:

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_1(t_1) & \mathbf{x}_2(t_1) & \cdots & \mathbf{x}_n(t_1) \\ \mathbf{x}_1(t_2) & \mathbf{x}_2(t_2) & \cdots & \mathbf{x}_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1(t_T - 1) & \mathbf{x}_2(t_T - 1) & \cdots & \mathbf{x}_n(t_T - 1) \end{pmatrix} \tag{7.2}$$

$$y_i = [y_i(t_2), y_i(t_3), ..., y_i(t_T)]^\top, \tag{7.3}$$

where T is the last time point and n is the number of parental genes. This vector and matrix are used in the regression analysis to determine the covariance hyperparameters of the Gaussian process. In the event that the parental set is not known, CSI can model the function by taking each possible parental set in turn. For example, in a network of 3 genes, the possible parental sets for gene 1 are: {1};{2};{3};{1,2};{1,3};{2,3};{1,2,3}.

The parental sets are then assigned a likelihood, called the 'joint distribution', by marginalising over $f(x)$.

This regression method also has two ways of calculating the marginal likelihood and hyperparameters - using expectation maximisation or Markov chain Monte Carlo (MCMC).

7.1.2 GENIE3

The expression of gene j in the kth experiment can be written as:

$$x_k^j = f_j(\mathbf{x}_k^{-j}) + \epsilon_k, \forall k, \tag{7.4}$$

where $\mathbf{x}_k^{-j}$ is a vector containing the expression values of all the transcription factors except j in the kth experiment, function f_j is non-linear and can involve the combinatorial expression of several genes, and ϵ_k is random noise with zero mean. A local ranking of all the genes apart from j is performed by minimising the square error loss function using the Random Forest or Extra-Trees method, with the final aim of reducing the variance of the target gene:

$$\sum_{k=1}^{N} (x_k^j - f_j(\mathbf{x}_k^{-j}))^2 \tag{7.5}$$

Each tree provides a ranking of the transcription factors and their ability to explain the variance in the output variable, which is normalised and averaged.

7.1.3 GRENITS

The gene expression of gene $g = 1, ..., G$ measured at time $t = 1, ..., T$ is modelled as a linear autoregressive process thus:

$$y_g^{t+1} = \mu_g + \sum_{j=1}^{G} y_j^t \tilde{\beta}_{jg} + \varepsilon_g^t, \tag{7.6}$$

where μ_g is the basal expression of gene g, $\tilde{\beta}_{jg} = \gamma_{jg}\beta_{jg}$ and measures the influence of gene j on gene g, with $\beta_{jg} \in \mathbb{R}$ and $\gamma_{jg} = 1$ if j regulates g and 0 otherwise. ε_g^t is an error term centred at zero and usually Gaussian.

Due to the nature of noisy measurements, x_{gr}^t is observed instead of the true gene expression y_g^t, such that:

$$x_{gr}^t = y_g^t + \eta_{gr}^t, \qquad r = 1, ..., R, \tag{7.7}$$

where η_{gr}^t is the measurement error term with zero mean.

Equations (7.6) and (7.7) are combined to form the likelihood and priors are specified over all the unknowns. As the posterior distribution cannot be analytically

determined, an MCMC algorithm is applied with Gibbs sampling for the parameters, allowing sampling from the posterior distribution.

7.1.4 Inferelator

tlCLR uses ODEs to model gene expression from a time series, and calculates the mutual information between all pairs of genes and finally applies a filter to remove interactions which are unlikely. Mutual information describes the dependency between two random variables X and Y as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}, \tag{7.8}$$

where $p(x,y)$ is the joint probability distribution function of X and Y , and $p(x)$ and $p(y)$ are the marginal probabilities that $X = x$ and $Y = y$, respectively. This makes it a symmetric measure (ie $I(X;Y) = I(Y;X)$), which is the basis for CLR. Directionality is an additional advantage to tlCLR and is achieved by using dynamic-MI instead of the static-MI described above. Two Z scores are then calculated based on the static and dynamic-MI, giving a final tlCLR score.

Rankings of potential regulators of x_i obtained from tlCLR are used by Inferelator to learn a sparse dynamical model of regulation for the target according to:

$$\frac{dx_i(t)}{dt} = -\alpha_i x_i + \sum_{j=1}^{P} \beta_{i,j} x_j^i(t), \qquad i = 1, ..., N \tag{7.9}$$

where x^i is the subset of regulators inferred previously from tlCLR, P is the number of regulators in that subset, $\alpha_i > 0$ is the first-order degradation rate of x_i and $\beta_{i,j}$ are a set of dynamical parameters to be estimated. The value of $\beta_{i,j}$ describes the extent of the regulation of target gene x_i by regulator x_j^i.

LARS then parameterises the ODE describing the temporal evolution of x_i in (7.9) while constraining β and preventing overfitting.

7.1.5 TIGRESS

The aim of the algorithm is to infer a score $s : \mathcal{E} \to \mathbb{R}$ for each potential regulation pair $(t,g) \in \mathcal{E}$, where g belongs to the set of all p genes $\mathcal{G} = [1,p]$ and t is in the transcription factor subset $\mathcal{T} \subset [1,p]$.

Similarly to the other algorithms, TIGRESS considers the expression of gene g as X_g, which is a linear or non-linear function of the transcription factors acting on g, $f_g(X_{\mathcal{T}_g})$:

$$X_g = f_g(X_{\mathcal{T}_g}) + \epsilon, \tag{7.10}$$

with ϵ being some noise and $X_{\mathcal{T}_g} = \{X_t, t \in \mathcal{T}_g\}$ the set of transcription factors that can regulate gene g. Rather than inferring $f_g(X_{\mathcal{T}_g})$, the aim is to score the probability of each transcription factor being involved in regulating g:

$$f_g(X_{\mathcal{T}_g}) = \sum_{t \in \mathcal{T}} \beta_{t,g} X_t, \tag{7.11}$$

such that the score $s_g(t)$ will asses the probability that $\beta_{t,g}$ is non-zero.

LARS is used for scoring; it models the regression function linearly by iteratively adding transcription factors in the model in order to get a better prediction for X_g. This provides a ranking of the regulators tested, but not the desired score, which is obtained after performing stability selection. Stability selection involves running LARS many times on data that has been randomly perturbed and scoring each transcription factor by the number of times it gets selected as a top feature after L steps. The perturbation consists of splitting the data into two halves, with the levels of the transcription factors multiplied by a random number $[\alpha, 1]$ for some $\alpha \leq 1$. LARS is performed on these modified datasets. The score for each (t, g) pair is the area under each curve up to L steps, which is then normalised per target gene:

$$\sum_t s_{area}(t, g) = \frac{1}{L} \sum_{l=1}^{L} F(g, t, L) \tag{7.12}$$

7.2 Appendix B - Covariance kernels for Gaussian process regression

7.2.1 Squared exponential kernel

The simple isotropic squared exponential (SE) covariance function is shown here:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 exp\left[\frac{-(\mathbf{x} - \mathbf{x}')^2}{2l^2}\right], \tag{7.13}$$

where the hyperparameters are: l, the characteristic length scale and σ_f, the standard deviation that describes the covariance function.

The SE function with automatic relevance determination learns an individual length scale hyperparameter for each input (regulator).

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 exp\left[-\frac{1}{2} \sum_{d=1}^{D} \left(\frac{x_d - x'_d}{l_d}\right)^2\right], \tag{7.14}$$

where d represents each input (regulator) and l_d is the length scale hyperparameter for each input.

7.2.2 Rational quadratic kernel

The rational quadratic (RQ) kernel is the equivalent of adding SE kernels together of different length scales. As a result, GP priors with this covariance function result in functions that vary smoothly between length scales.

$$k(\mathbf{x}, \mathbf{x}')) = \sigma_f^2 \left[1 + \frac{(\mathbf{x} - \mathbf{x}')^2}{2\alpha l^2}\right]^{-\alpha}, \tag{7.15}$$

where α is a scaling parameter for the different length scales.

Similarly to equation 7.14, the RQ kernel with ARD consists of individual length scale hyperparameters for each input (regulator).

7.2.3 Linear kernel

Using a linear kernel is the equivalent of doing Bayesian linear regression. The equation for this is:

$$k(\mathbf{x}, \mathbf{x}')) = \sigma_a^2 + \mathbf{x} \cdot \mathbf{x}' \tag{7.16}$$

7.2.4 Kernel operations

Kernels can be combined by adding together two independent kernels:

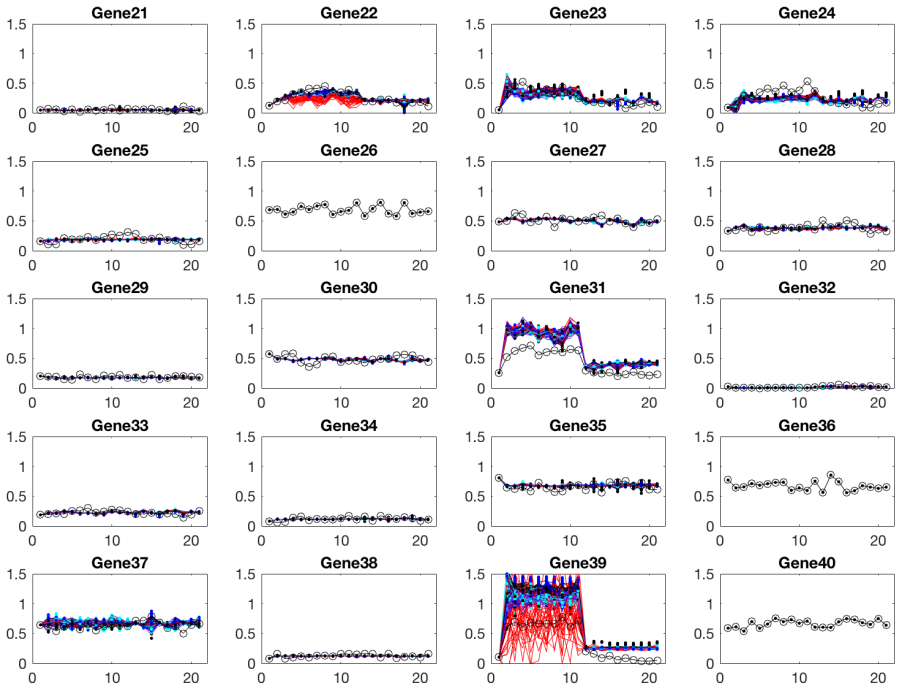
$$k(\mathbf{x}, \mathbf{x}')) = k_1(\mathbf{x}, \mathbf{x}')) + k_2(\mathbf{x}, \mathbf{x}')) \tag{7.17}$$

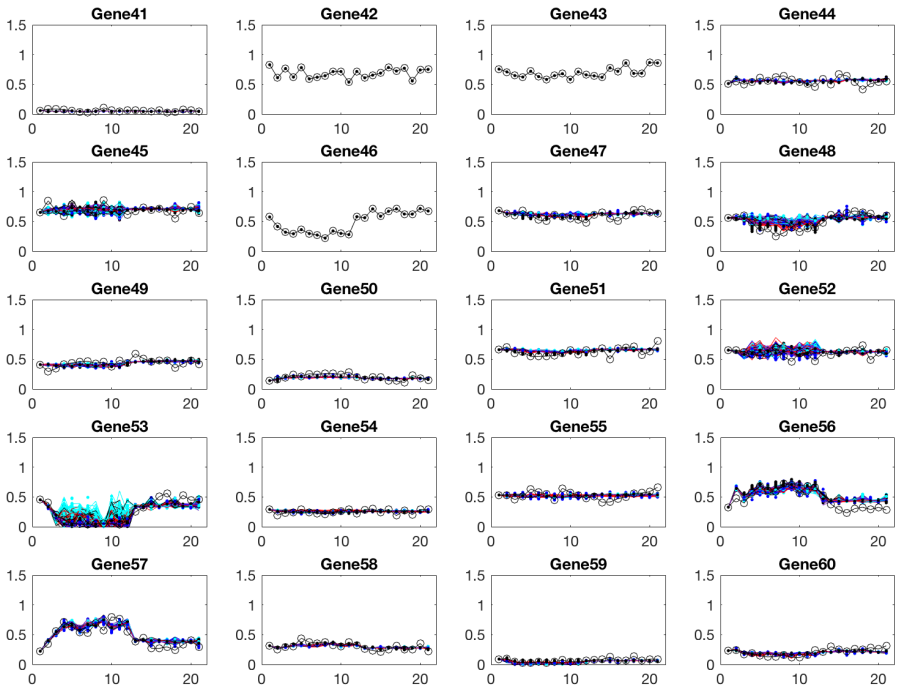
or obtaining the product of two independent kernels:

$$k(\mathbf{x}, \mathbf{x}')) = k_1(\mathbf{x}, \mathbf{x}'))k_2(\mathbf{x}, \mathbf{x}')) \tag{7.18}$$

7.3 Appendix C - GPDS model gene simulations

7.3.1 Simulations of validation WT DREAM100 network




Gene41
Gene42
Gene43
Gene44
Gene45
Gene46
Gene47
Gene48
Gene49
Gene50
Gene51
Gene52
Gene53
Gene54
Gene55
Gene56
Gene57
Gene58
Gene59
Gene60

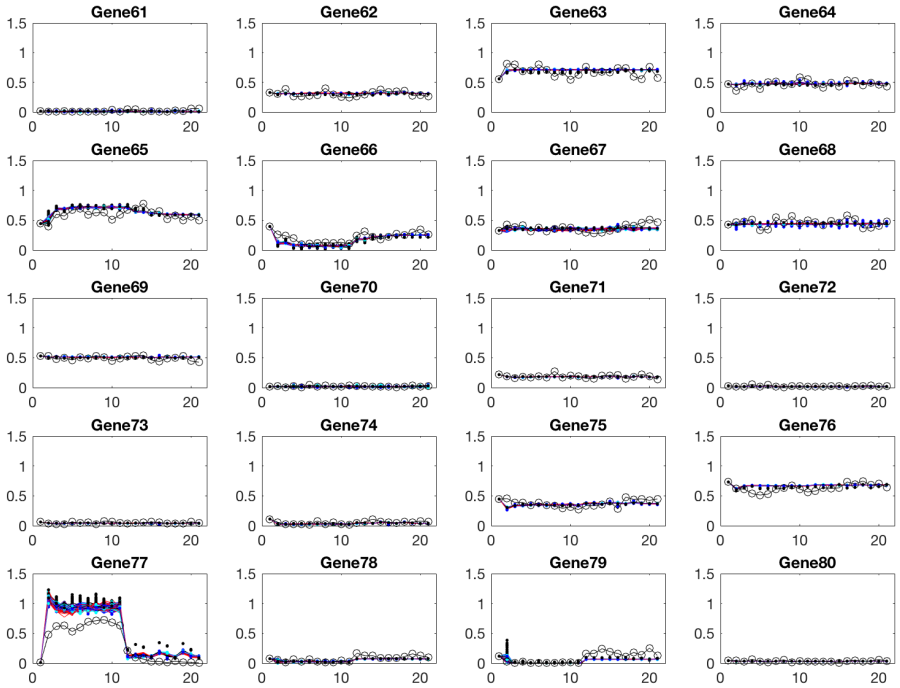

Gene61
Gene62
Gene63
Gene64
Gene65
Gene66
Gene67
Gene68
Gene69
Gene70
Gene71
Gene72
Gene73
Gene74
Gene75
Gene76
Gene77
Gene78
Gene79
Gene80

Figure 7.1: *(previous page)*: GPDS models learn the functions connecting regulators and targets from training data and made to predict gene expression data for the validation dataset - which has the same underlying network but different perturbations applied to some of the genes. Gene expression data from the validation dataset is plotted in black with round markers, and simulations made using the 8 covariance functions defined above are also plotted. All models were trained on datasets with 10 different perturbations and 1 biological repeat. Simulations were started from the second time point onwards. Some gene levels is fixed so its levels are not simulated - more details are given in Section 4.4.1.3.

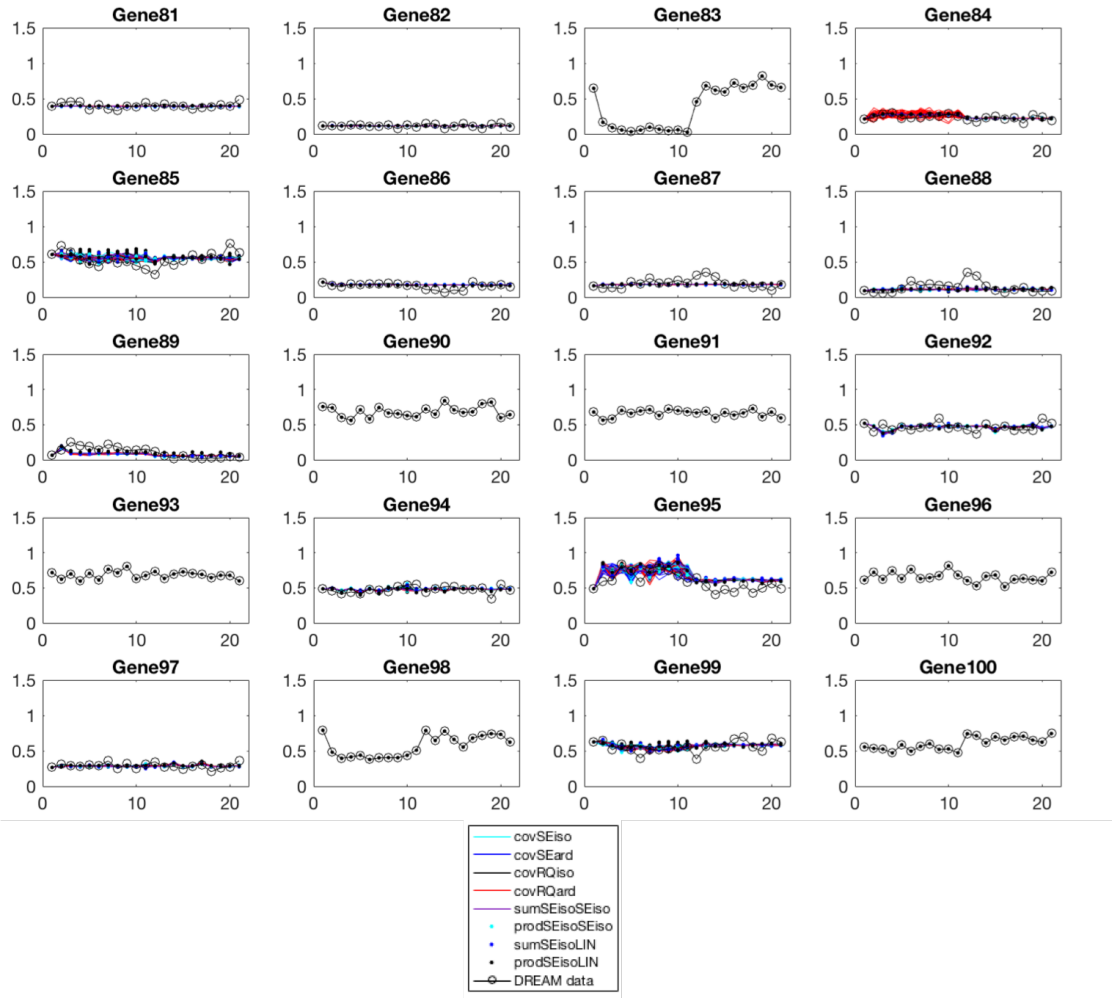
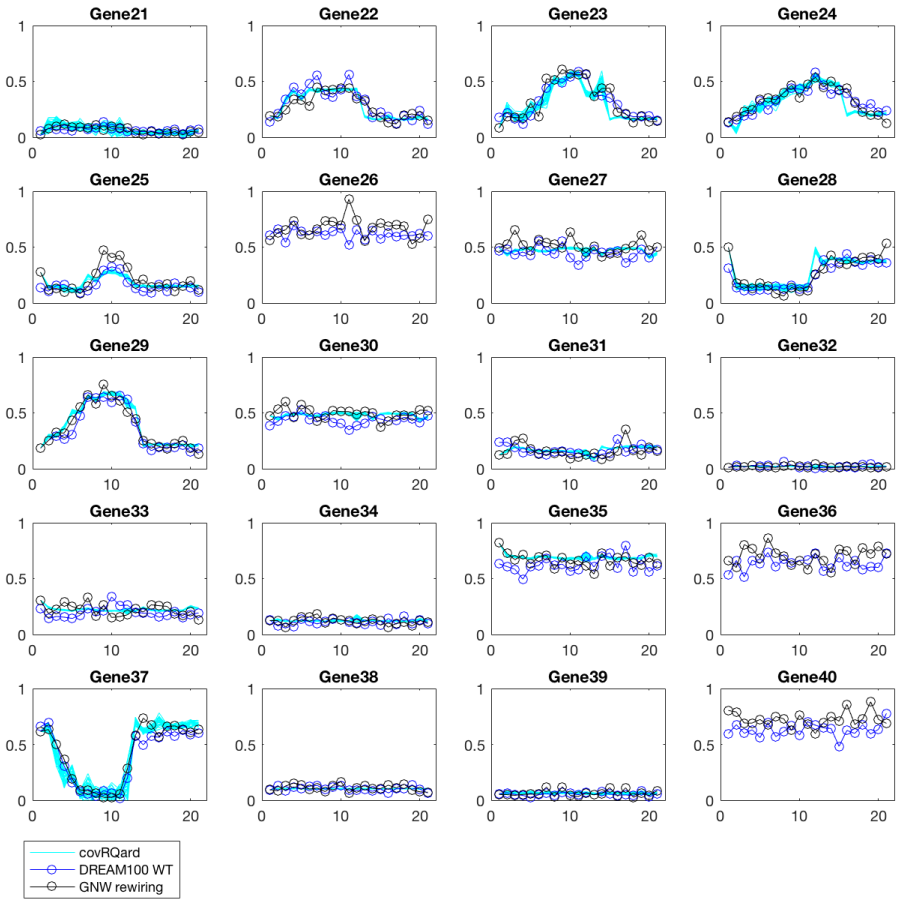
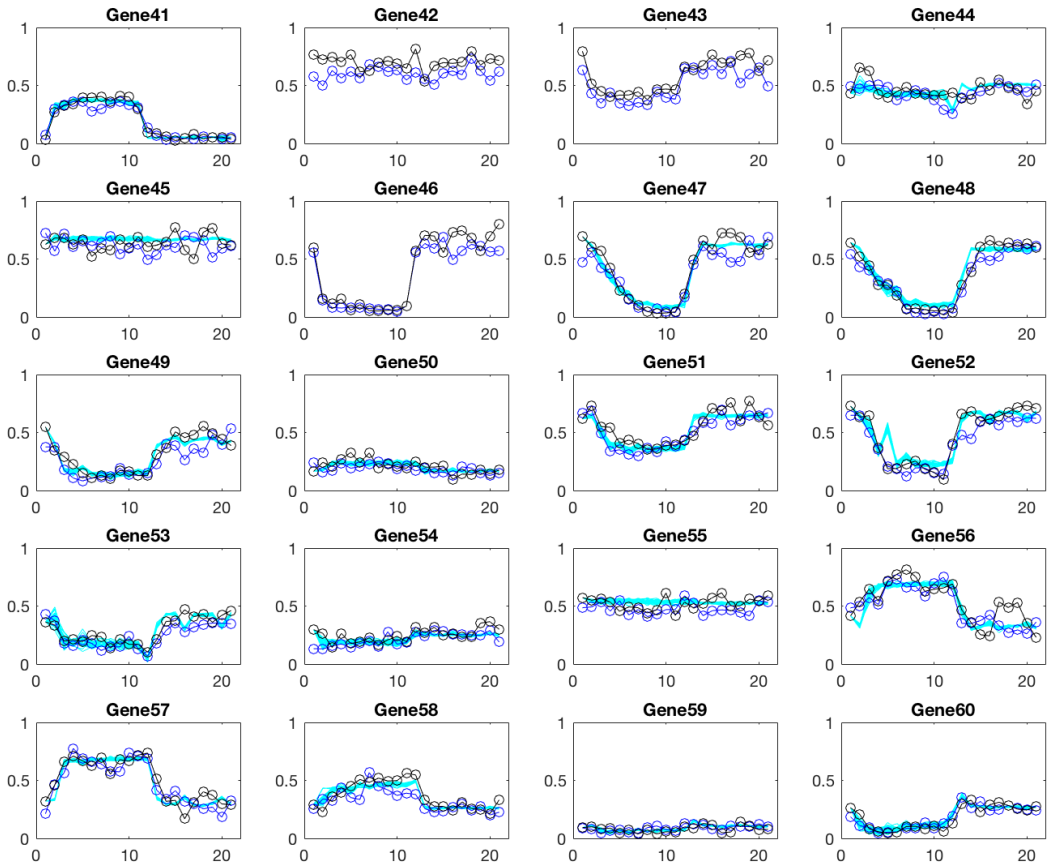
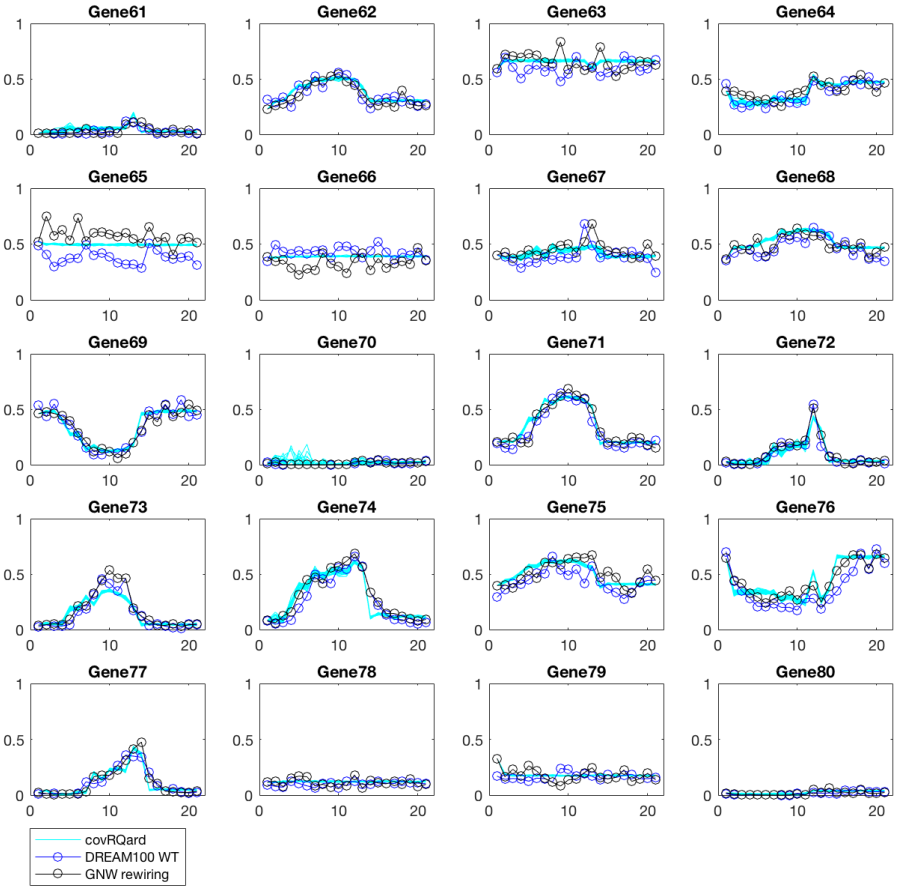


Figure 7.1: GPDS model simulations of the 8 covariance functions for genes from the validation DREAM100 dataset.

7.3.2 Simulations of rewired DREAM100 network




Gene41
Gene42
Gene43
Gene44
Gene45
Gene46
Gene47
Gene48
Gene49
Gene50
Gene51
Gene52
Gene53
Gene54
Gene55
Gene56
Gene57
Gene58
Gene59
Gene60


Gene61
Gene62
Gene63
Gene64
Gene65
Gene66
Gene67
Gene68
Gene69
Gene70
Gene71
Gene72
Gene73
Gene74
Gene75
Gene76
Gene77
Gene78
Gene79
Gene80
covRQard
DREAM100 WT
GNW rewiring

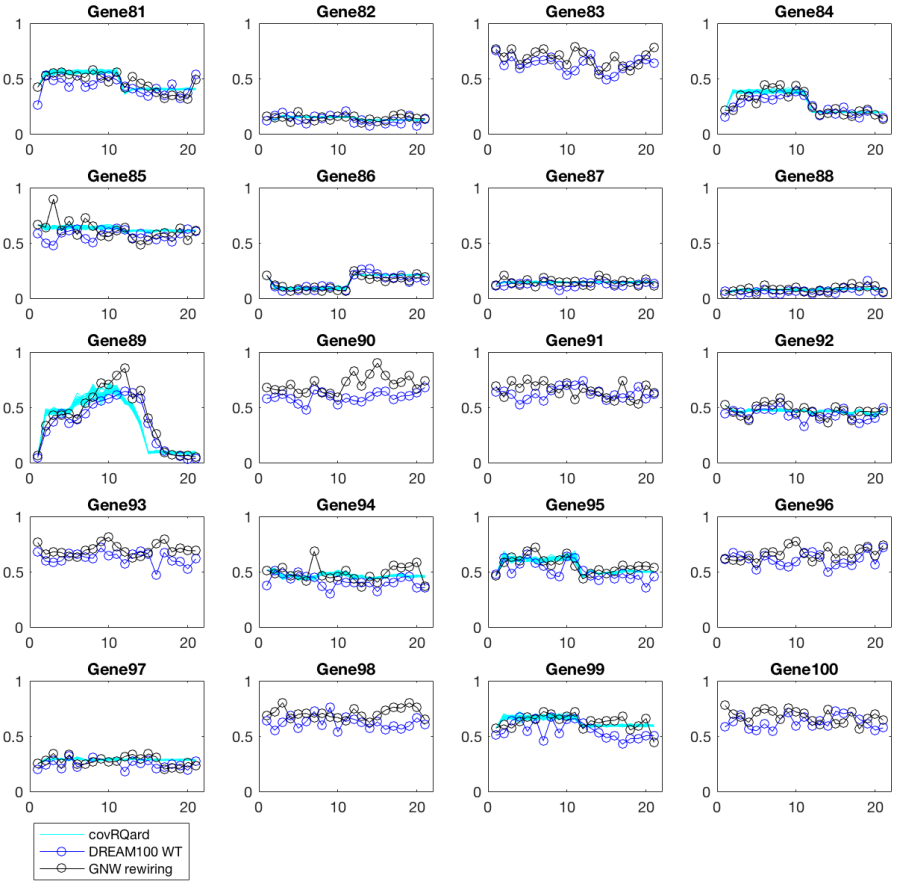


Figure 7.2: GPDS model rewiring simulations of DREAM100 network gene expressions obtained by replacing regulators of G5 with the regulators of G1. Solid black line with circular markers: the gene expression data of the rewired network as given by GNW. Solid blue line with circular markers: the gene expression of data from the WT network with the same perturbation as the rewired network. Cyan solid lines are 30 simulations from the covRQard covariance kernel for each gene. Simulations are started off at the second time point and some genes in the DREAM100 network are fixed as outlined in Section 4.4.2.3.

Chapter 8

Bibliography

Abe, H., Urao, T., Ito, T., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2003). Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *The Plant Cell*, **15**(1), 63–78.

AbuQamar, S., Moustafa, K., and Tran, L. S. (2017). Mechanisms and strategies of plant defense against Botrytis cinerea. *Critical Reviews in Biotechnology*, **37**(2), 262–274.

Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., *et al.* (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, **14**(11), 1083.

Äijö, T. and Lähdesmäki, H. (2009). Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, **25**(22), 2937–2944.

Aldana, M., Balleza, E., Kauffman, S., and Resendiz, O. (2007). Robustness and evolvability in genetic regulatory networks. *Journal of Theoretical Biology*, **245**(3), 433–448.

Ali, G. S., Prasad, K., Day, I., and Reddy, A. (2007). Ligand-dependent reduction in the membrane mobility of FLAGELLIN SENSITIVE2, an Arabidopsis receptor-like kinase. *Plant and Cell Physiology*, **48**(11), 1601–1611.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetic*, **8**(6), 450–461.

Alonso, J. M., Hirayama, T., Roman, G., Nourizadeh, S., and Ecker, J. R. (1999). EIN2, a bifunctional transducer of ethylene and stress responses in Arabidopsis. *Science*, **284**(5423), 2148–2152.

Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C. C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D. E., Marchand, T., Risseeuw, E., Brogden, D., Zeko, A., Crosby, W. L., Berry, C. C., and Ecker, J. R. (2003). Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, **301**(5633), 653–657.

Altun, Y., Hofmann, T., and Smola, A. J. (2004). Gaussian process classification for segmenting and annotating sequences. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 4. ACM.

Alves, M. S., Dadalto, S. P., Gonçalves, A. B., De Souza, G. B., Barros, V. A., and Fietto, L. G. (2013). Plant bZIP transcription factors responsive to pathogens: a review. *International Journal of Molecular Sciences*, **14**(4), 7815–7828.

Anderson, J. P., Badruzsaufari, E., Schenk, P. M., Manners, J. M., Desmond, O. J., Ehlert, C., Maclean, D. J., Ebert, P. R., and Kazan, K. (2004). Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in Arabidopsis. *The Plant Cell*, **16**(12), 3460–3479.

Andersson, M., Turesson, H., Nicolia, A., Fält, A.-S., Samuelsson, M., and Hofvander, P. (2017). Efficient targeted multiallelic mutagenesis in tetraploid potato (Solanum tuberosum) by transient CRISPR-Cas9 expression in protoplasts. *Plant Cell Reports*, **36**(1), 117–128.

Andrews, S. *et al.* (2010). FastQC: a quality control tool for high throughput sequence data. `http://www.Bioinformatics.babraham.ac.uk/projects/fastqc/`.

Ang, J. and McMillen, D. (2013). Physical constraints on biological integral control design for homeostasis and sensory adaptation. *Biophysical Journal*, **104**(2), 505–515.

Ang, J., Bagh, S., Ingalls, B. P., and McMillen, D. R. (2010). Considerations for using integral feedback control to construct a perfectly adapting synthetic gene network. *Journal of Theoretical Biology*, **266**(4), 723–738.

Antunes, M. S., Morey, K. J., Smith, J. J., Albrecht, K. D., Bowen, T. A., Zdunek, J. K., Troupe, J. F., Cuneo, M. J., Webb, C. T., Hellinga, H. W., *et al.* (2011). Programmable ligand detection system in plants through a synthetic signal transduction pathway. *PLoS One*, **6**(1), e16292.

Aoyama, T. and Chua, N.-H. (1997). A glucocorticoid-mediated transcriptional induction system in transgenic plants. *The Plant Journal*, **11**(3), 605–612.

Arrieta-Ortiz, M. L., Hafemeister, C., Bate, A. R., Chu, T., Greenfield, A., Shuster, B., Barry, S. N., Gallitto, M., Liu, B., Kacmarczyk, T., *et al.* (2015). An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. *Molecular Systems Biology*, **11**(11), 839.

Asai, T., Tena, G., Plotnikova, J., Willmann, M. R., Chiu, W.-L., Gomez-Gomez, L., Boller, T., Ausubel, F. M., and Sheen, J. (2002). MAP kinase signalling cascade in Arabidopsis innate immunity. *Nature*, **415**(6875), 977.

Atkinson, N. J. and Urwin, P. E. (2012). The interaction of plant biotic and abiotic stresses: from genes to the field. *Journal of Experimental Botany*, **63**(10), 3523–3543.

Audenaert, K., De Meyer, G. B., and Höfte, M. M. (2002). Abscisic acid determines basal susceptibility of tomato toBotrytis cinerea and suppresses salicylic acid-dependent signaling mechanisms. *Plant Physiology*, **128**(2), 491–501.

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME suite. *Nucleic Acids Research*, **43**(W1), W39–W49.

Balazadeh, S., Schildhauer, J., Araújo, W. L., Munné-Bosch, S., Fernie, A. R., Proost, S., Humbeck, K., and Mueller-Roeber, B. (2014). Reversal of senescence by N resupply to N-starved Arabidopsis thaliana: transcriptomic and metabolomic consequences. *Journal of Experimental Botany*, **65**(14), 3975–3992.

Band, L. R., Fozard, J. A., Godin, C., Jensen, O. E., Pridmore, T., Bennett, M. J., and King, J. R. (2012). Multiscale systems analysis of root growth and development: modeling beyond the network and cellular scales. *The Plant Cell*, **24**(10), 3892–3906.

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.-K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., *et al.* (2010). Rewiring of genetic networks in response to DNA damage. *Science*, **330**(6009), 1385–1389.

Banf, M. and Rhee, S. Y. (2017). Computational inference of gene regulatory networks: Approaches, limitations and opportunities. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1860**(1), 41–52.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular Systems Biology*, **3**(1).

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network Biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**(2), 101.

Bari, R. and Jones, J. D. (2009). Role of plant hormones in plant defence responses. *Plant Molecular Biology*, **69**(4), 473–488.

Barkai, N. and Leibler, S. (1997). Robustness in simple biochemical networks. *Nature*, **387**(6636), 913.

Barman, S. and Kwon, Y.-K. (2017). A novel mutual information-based Boolean network inference method from time-series gene expression data. *PLoS One*, **12**(2), e0171097.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**(2), 281–297.

Basu, S., Gerchman, Y., Collins, C. H., Arnold, F. H., and Weiss, R. (2005). A synthetic multicellular system for programmed pattern formation. *Nature*, **434**(7037), 1130–1134.

Battisti, D. S. and Naylor, R. L. (2009). Historical warnings of future food insecurity with unprecedented seasonal heat. *Science*, **323**(5911), 240–244.

Baumstark, R., Hänzelmann, S., Tsuru, S., Schaerli, Y., Francesconi, M., Mancuso, F. M., Castelo, R., and Isalan, M. (2015). The propagation of perturbations in rewired bacterial gene networks. *Nature Communications*, **6**, 10105.

Bayer, E. M., Smith, R. S., Mandel, T., Nakayama, N., Sauer, M., Prusinkiewicz, P., and Kuhlemeier, C. (2009). Integration of transport-based models for phyllotaxis and midvein formation. *Genes & Development*, **23**(3), 373–384.

Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. L. (2004). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, **21**(3), 349–356.

Bebber, D. P., Ramotowski, M. A., and Gurr, S. J. (2013). Crop pests and pathogens move polewards in a warming world. *Nature climate change*, **3**(11), 985.

Bebber, D. P., Holmes, T., and Gurr, S. J. (2014). The global spread of crop pests and pathogens. *Global Ecology and Biogeography*, **23**(12), 1398–1407.

Bechtold, U., Penfold, C. A., Jenkins, D. J., Legaie, R., Moore, J. D., Lawson, T., Matthews, J. S., Vialet-Chabrand, S. R., Baxter, L., Subramaniam, S., *et al.* (2016). Time-series transcriptomics reveals that AGAMOUS-LIKE22 links primary metabolism to developmental processes in drought-stressed Arabidopsis. *The Plant Cell*, pages TPC2015–00910.

Beintema, N., Elliott, H., *et al.* (2009). Setting meaningful investment targets in agricultural research and development: challenges, opportunities and fiscal realities. In *How to feed the World in 2050. Proceedings of a technical meeting of experts, Rome, Italy, 24-26 June 2009*, pages 1–29. Food and Agriculture Organization of the United Nations (FAO).

Ben-Gal, I. (2008). Bayesian networks. In *Encyclopedia of Statistics in Quality and Reliability*, volume 1. Wiley Online Library.

Benschop, J. J., Mohammed, S., O'Flaherty, M., Heck, A. J., Slijper, M., and Menke, F. L. (2007). Quantitative phosphoproteomics of early elicitor signaling in Arabidopsis. *Molecular & Cellular Proteomics*, **6**(7), 1198–1214.

Berrocal-Lobo, M., Molina, A., and Solano, R. (2002). Constitutive expression of ETHYLENE-RESPONSE-FACTOR1 in Arabidopsis confers resistance to several necrotrophic fungi. *The Plant Journal*, **29**(1), 23–32.

Besseau, S., Li, J., and Palva, E. T. (2012). WRKY54 and WRKY70 co-operate as negative regulators of leaf senescence in Arabidopsis thaliana. *Journal of Experimental Botany*, **63**(7), 2667–2679.

Bhardwaj, N., Kim, P. M., and Gerstein, M. B. (2010). Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci. Signal.*, **3**(146), ra79–ra79.

Bhatnagar-Mathur, P., Vadez, V., and Sharma, K. K. (2008). Transgenic approaches for abiotic stress tolerance in plants: retrospect and prospects. *Plant Cell Reports*, **27**(3), 411–424.

Bilgin, D. D., Zavala, J. A., Zhu, J., Clough, S. J., Ort, D. R., and DeLUCIA, E. H. (2010). Biotic stress globally downregulates photosynthesis genes. *Plant, Cell & environment*, **33**(10), 1597–1613.

Birkenbihl, R. P., Diezel, C., and Somssich, I. E. (2012). Arabidopsis WRKY33 is a key transcriptional regulator of hormonal and metabolic responses towards Botrytis cinerea infection. *Plant Physiology*, pages pp–111.

Birnbaum, K., Jung, J. W., Wang, J. Y., Lambert, G. M., Hirst, J. A., Galbraith, D. W., and Benfey, P. N. (2005). Cell type–specific expression profiling in plants via cell sorting of protoplasts from fluorescent reporter lines. *Nature Methods*, **2**(8), 615.

Blanc, G. and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant Cell*, **16**(7), 1679–1691.

Bleecker, A. B. and Kende, H. (2000). Ethylene: a gaseous signal molecule in plants. *Annual review of Cell and Developmental Biology*, **16**(1), 1–18.

Boehm, R. (2007). Bioproduction of therapeutic proteins in the 21st century and the role of plants and plant cells as production platforms. *Annals of the New York Academy of Sciences*, **1102**(1), 121–134.

Boisen, A., Christensen, T., Fu, W., Gorbanev, Y., Hansen, T., Jensen, J., Klitgaard, S., Pedersen, S., Riisager, A., Ståhlberg, T., *et al.* (2009). Process integration for the conversion of glucose to 2, 5-furandicarboxylic acid. *Chemical Engineering Research and Design*, **87**(9), 1318–1327.

Boller, T. and Felix, G. (2009). A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annual review of Plant Biology*, **60**, 379–406.

Boller, T. and He, S. Y. (2009). Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science*, **324**(5928), 742–744.

Bonaventure, G., Gfeller, A., Rodríguez, V. M., Armand, F., and Farmer, E. E. (2007). The fou2 gain-of-function allele and the wild-type allele of Two Pore Channel 1 contribute to different extents or by different mechanisms to defense gene expression in Arabidopsis. *Plant and Cell Physiology*, **48**(12), 1775–1789.

Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., Baliga, N. S., and Thorsson, V. (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology*, **7**(5), 1.

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., Wang, L. Y., Gerstein, M., and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. *Science*, **317**(5839), 815–819.

Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., and Zhu, J.-K. (2005). Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell*, **123**(7), 1279–1291.

Boudsocq, M., Willmann, M. R., McCormack, M., Lee, H., Shan, L., He, P., Bush, J., Cheng, S.-H., and Sheen, J. (2010). Differential innate immune signalling via Ca 2+ sensor protein kinases. *Nature*, **464**(7287), 418.

Bourgin, J.-P., Chupeau, Y., and Missonier, C. (1979). Plant regeneration from mesophyll protoplasts of several Nicotiana species. *Physiologia Plantarum*, **45**(2), 288–292.

Boyer, J. S. (1982). Plant productivity and environment. *Science*, **218**(4571), 443–448.

Bray, E. A. (2000). Response to abiotic stress. *Biochemistry and Molecular Biology of Plants*, pages 1158–1203.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**(5), 525.

Breeze, E., Harrison, E., McHattie, S., Hughes, L., Hickman, R., Hill, C., Kiddle, S., Kim, Y.-s., Penfold, C. A., Jenkins, D., Zhang, C., Morris, K., Jenner, C., Jackson, S., Thomas, B., Tabrett, A., Legaie, R., Moore, J. D., Wild, D. L., Ott, S., Rand, D., Beynon, J., Denby, K., Mead, A., and Buchanan-Wollaston, V. (2011). High-resolution temporal profiling of transcripts during Arabidopsis leaf senescence reveals a distinct chronology of processes and regulation. *The Plant Cell*, **23**(3), 873–894.

Briat, C., Gupta, A., and Khammash, M. (2016a). Antithetic integral feedback ensures robust perfect adaptation in noisy biomolecular networks. *Cell systems*, **2**(1), 15–26.

Briat, C., Zechner, C., and Khammash, M. (2016b). Design of a synthetic integral

feedback circuit: dynamic analysis and DNA implementation. *ACS Synthetic Biology*, **5**(10), 1108–1116.

Bright, J., Desikan, R., Hancock, J. T., Weir, I. S., and Neill, S. J. (2006). ABA-induced NO generation and stomatal closure in Arabidopsis are dependent on H2O2 synthesis. *The Plant Journal*, **45**(1), 113–122.

Brkljacic, J. and Grotewold, E. (2017). Combinatorial control of plant gene expression. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1860**(1), 31–40.

Broekaert, W. F., Delauré, S. L., De Bolle, M. F., and Cammue, B. P. (2006). The role of ethylene in host-pathogen interactions. *Annual Review of Phytopathology*, **44**, 393–416.

Bruggeman, F. J. and Westerhoff, H. V. (2007). The nature of systems biology. *Trends in Microbiology*, **15**(1), 45–50.

Brutus, A., Sicilia, F., Macone, A., Cervone, F., and De Lorenzo, G. (2010). A domain swap approach reveals a role of the plant wall-associated kinase 1 (WAK1) as a receptor of oligogalacturonides. *Proceedings of the National Academy of Sciences*, **107**(20), 9452–9457.

Buchanan, B. B., Gruissem, W., and Jones, R. L. (2015). Responses to Abiotic Stresses. In *Biochemistry and Molecular Biology of Plants*. John Wiley & Sons.

Buchanan-Wollaston, V., Wilson, Z., Tardieu, F., Beynon, J., and Denby, K. (2017). Harnessing diversity from ecosystems to crops to genes. *Food and Energy Security*, **6**(1), 19–25.

Bueso, E., Ibañez, C., Sayas, E., Muñoz-Bertomeu, J., Gonzalez-Guzmán, M., Rodriguez, P. L., and Serrano, R. (2014). A forward genetic approach in Arabidopsis thaliana identifies a RING-type ubiquitin ligase as a novel determinant of seed longevity. *Plant Science*, **215**, 110–116.

Bujdoso, N. and Davis, S. J. (2013). Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of Arabidopsis thaliana. *Frontiers in Plant Science*, **4**, 3.

Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., Blake, J. A., and Group, M. G. D. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Research*, **36**(suppl_1), D724–D728.

Butte, A. J. and Kohane, I. S. (1999). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific.

Caplan, J., Padmanabhan, M., and Dinesh-Kumar, S. P. (2008). Plant NB-LRR immune receptors: from recognition to transcriptional reprogramming. *Cell Host & Microbe*, **3**(3), 126–135.

Cappuccio, A., Tieri, P., and Castiglione, F. (2015). Multiscale modelling in immunology: a review. *Briefings in Bioinformatics*, **17**(3), 408–418.

Cardinale, S., Joachimiak, M. P., and Arkin, A. P. (2013). Effects of genetic variation on the E. coli host-circuit interface. *Cell Reports*, **4**(2), 231–237.

Carere, C. R., Sparling, R., Cicek, N., and Levin, D. B. (2008). Third generation biofuels

via direct cellulose fermentation. *International Journal of Molecular Sciences*, **9**(7), 1342–1360.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**(13), 2933–2942.

Century, K., Reuber, T. L., and Ratcliffe, O. J. (2008). Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products. *Plant Physiology*, **147**(1), 20–29.

Ceroni, F., Boo, A., Furini, S., Gorochowski, T. E., Borkowski, O., Ladak, Y. N., Awan, A. R., Gilbert, C., Stan, G.-B., and Ellis, T. (2018). Burden-driven feedback control of gene expression. *Nature Methods*, **15**(5), 387.

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., and Cercignani, M. (2015). Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *Neuroimage*, **112**, 232–243.

Chang, C., Kwok, S. F., Bleecker, A. B., and Meyerowitz, E. M. (1993). Arabidopsis ethylene-response gene ETR1: similarity of product to two-component regulators. *Science*, **262**(5133), 539–544.

Chen, T., Morris, J., and Martin, E. (2007). Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems*, **87**(1), 59–71.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., *et al.* (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**(6), 1106–1117.

Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., Zhang, Y., Kim, T.-K., He, H. H., Zieba, J., *et al.* (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, **9**(6), 609.

Cheng, C.-Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *The Plant Journal*, **89**(4), 789–804.

Cheng, F., Yu, J., and Xiong, H. (2010). Facial expression recognition in JAFFE dataset based on Gaussian process classification. *IEEE Transactions on Neural Networks*, **21**(10), 1685–1690.

Chenu, K., Chapman, S. C., Tardieu, F., McLean, G., Welcker, C., and Hammer, G. L. (2009). Simulating the Yield Impacts of Organ-Level Quantitative Trait Loci Associated with Drought Response in Maize-A'Gene-to-Phenotype'Modeling Approach. *Genetics*.

Cheong, Y. H., Chang, H.-S., Gupta, R., Wang, X., Zhu, T., and Luan, S. (2002). Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in Arabidopsis. *Plant Physiology*, **129**(2), 661–677.

Chien, P.-J., Sheu, F., and Lin, H.-R. (2007). Coating citrus (Murcott tangor) fruit with low molecular weight chitosan increases postharvest quality and shelf life. *Food Chemistry*, **100**(3), 1160–1164.

Chini, A., Fonseca, S., Fernandez, G., Adie, B., Chico, J., Lorenzo, O., Garcia-Casado, G., Lopez-Vidriero, I., Lozano, F., Ponce, M., *et al.* (2007). The JAZ family of repressors is the missing link in jasmonate signalling. *Nature*, **448**(7154), 666.

Choe, S., Fujioka, S., Noguchi, T., Takatsuto, S., Yoshida, S., and Feldmann, K. A. (2001). Overexpression of DWARF4 in the brassinosteroid biosynthetic pathway results in increased vegetative growth and seed yield in Arabidopsis. *The Plant Journal*, **26**(6), 573–582.

Choi, H. W. and Klessig, D. F. (2016). DAMPs, MAMPs, and NAMPs in plant innate immunity. *BMC Plant Biology*, **16**(1), 232.

Chung, H. S. and Howe, G. A. (2009). A critical role for the TIFY motif in repression of jasmonate signaling by a stabilized splice variant of the JASMONATE ZIM-domain protein JAZ10 in Arabidopsis. *The Plant Cell*, **21**(1), 131–145.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A., *et al.* (2005). Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, **437**(7058), 529.

Citorik, R. J., Mimee, M., and Lu, T. K. (2014). Bacteriophage-based synthetic biology for the study of infectious diseases. *Current Opinion in Microbiology*, **19**, 59–69.

Clough, S. J. and Bent, A. F. (1998). Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *The plant journal*, **16**(6), 735–743.

Coego, A., Ramirez, V., Gil, M. J., Flors, V., Mauch-Mani, B., and Vera, P. (2005). An Arabidopsis homeodomain transcription factor, OVEREXPRESSOR OF CATIONIC PEROXIDASE 3, mediates resistance to infection by necrotrophic pathogens. *The Plant Cell*, **17**(7), 2123–2137.

Cong, B., Liu, J., and Tanksley, S. D. (2002). Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations. *Proceedings of the National Academy of Sciences*, **99**(21), 13606–13611.

Consortium, T. G. *et al.* (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**(7400), 635.

Contento, A. L., Kim, S.-J., and Bassham, D. C. (2004). Transcriptome profiling of the response of Arabidopsis suspension culture cells to Suc starvation. *Plant Physiology*, **135**(4), 2330–2347.

Covington, M. F., Maloof, J. N., Straume, M., Kay, S. A., and Harmer, S. L. (2008). Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biology*, **9**(8), R130.

Craufurd, P. and Peacock, J. (1993). Effect of heat and drought stress on sorghum (Sorghum bicolor). II. Grain yield. *Experimental Agriculture*, **29**(1), 77–86.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution*, **25**(7), 410–418.

Cushman, J. C. and Bohnert, H. J. (2000). Genomic approaches to plant stress tolerance. *Current Opinion in Plant Biology*, **3**(2), 117–124.

Dalchau, N., Baek, S. J., Briggs, H. M., Robertson, F. C., Dodd, A. N., Gardner, M. J., Stancombe, M. A., Haydon, M. J., Stan, G.-B., Gonçalves, J. M., and Webb, A. A. R. (2011). The circadian oscillator gene GIGANTEA mediates a long-term response of the Arabidopsis thaliana circadian clock to sucrose. *Proceedings of the National Academy of Sciences*, **108**(12), 5104–5109.

Dangl, J. L. and Jones, J. D. (2001). Plant pathogens and integrated defence responses to infection. *Nature*, **411**(6839), 826.

Darlington, A. P., Kim, J., Jiménez, J. I., and Bates, D. G. (2018). Dynamic allocation of orthogonal ribosomes facilitates uncoupling of co-expressed genes. *Nature Communications*, **9**(1), 695.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM.

De Clercq, I., Vermeirssen, V., Van Aken, O., Vandepoele, K., Murcha, M. W., Law, S. R., Inzé, A., Ng, S., Ivanova, A., Rombaut, D., *et al.* (2013). The membrane-bound NAC transcription factor ANAC013 functions in mitochondrial retrograde regulation of the oxidative stress response in Arabidopsis. *The Plant Cell*, **25**(9), 3472–3490.

De Smet, R. and Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology*, **15**(2), 168–176.

Dean, R., Van Kan, J. A. L., Pretorius, Z. A., Hammond-Kosack, K. E., Di Pietro, A., Spanu, P. D., Rudd, J. J., Dickman, M., Kahmann, R., Ellis, J., and Foster, G. D. (2012). The top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology*, **13**(4), 414–430.

Denby, K. J., Kumar, P., and Kliebenstein, D. J. (2004). Identification of Botrytis cinerea susceptibility loci in Arabidopsis thaliana. *The Plant Journal*, **38**(3), 473–486.

Denoux, C., Galletti, R., Mammarella, N., Gopalan, S., Werck, D., De Lorenzo, G., Ferrari, S., Ausubel, F. M., and Dewdney, J. (2008). Activation of defense response pathways by OGs and Flg22 elicitors in Arabidopsis seedlings. *Molecular Plant*, **1**(3), 423–445.

Desai, J. S., Sartor, R. C., Lawas, L. M., Jagadish, S. K., and Doherty, C. J. (2017). Improving Gene Regulatory Network Inference by Incorporating Rates of Transcriptional Changes. *Scientific reports*, **7**(1), 17244.

di Bernardo, D., Thompson, M. J., Gardner, T. S., Chobot, S. E., Eastwood, E. L., Wojtovich, A. P., Elliott, S. J., Schaus, S. E., and Collins, J. J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnology*, **23**(3), 377–383.

Diaz, R. J. and Rosenberg, R. (2008). Spreading dead zones and consequences for marine ecosystems. *science*, **321**(5891), 926–929.

Dobritsa, A. A., Geanconteri, A., Shrestha, J., Carlson, A., Kooyers, N., Coerper, D., Urbanczyk-Wochniak, E., Bench, B. J., Sumner, L. W., Swanson, R., *et al.* (2011). A large-scale genetic screen in Arabidopsis thaliana to identify genes involved in pollen exine production. *Plant Physiology*, pages pp–111.

Dodds, P. N. and Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant–pathogen interactions. *Nature Reviews Genetics*, **11**(8), 539.

Doebley, J. F., Gaut, B. S., and Smith, B. D. (2006). The molecular genetics of crop domestication. *Cell*, **127**(7), 1309–1321.

Dombrecht, B., Xue, G. P., Sprague, S. J., Kirkegaard, J. A., Ross, J. J., Reid, J. B., Fitt, G. P., Sewelam, N., Schenk, P. M., Manners, J. M., *et al.* (2007). MYC2 differentially modulates diverse jasmonate-dependent functions in Arabidopsis. *The Plant Cell*, **19**(7), 2225–2245.

Dong, C.-H., Hu, X., Tang, W., Zheng, X., Kim, Y. S., Lee, B.-h., and Zhu, J.-K. (2006). A putative Arabidopsis nucleoporin, AtNUP160, is critical for RNA export and required for plant tolerance to cold stress. *Molecular and Cellular Biology*, **26**(24), 9533–9543.

Drengstig, T., Ni, X. Y., Thorsen, K., Jolma, I. W., and Ruoff, P. (2012). Robust adaptation and homeostasis by autocatalysis. *The Journal of Physical Chemistry B*, **116**(18), 5355–5363.

Duan, C.-G., Wang, X., Xie, S., Pan, L., Miki, D., Tang, K., Hsu, C.-C., Lei, M., Zhong, Y., Hou, Y.-J., *et al.* (2017). A pair of transposon-derived proteins function in a histone acetyltransferase complex for active DNA demethylation. *Cell Research*, **27**(2), 226.

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., and Lepiniec, L. (2010). MYB transcription factors in Arabidopsis. *Trends in Plant Science*, **15**(10), 573–581.

Durrant, W. E. and Dong, X. (2004). Systemic acquired resistance. *Annual Review of Phytopathology*, **42**, 185–209.

Earley, K. W., Haag, J. R., Pontes, O., Opper, K., Juehne, T., Song, K., and Pikaard, C. S. (2006). Gateway-compatible vectors for plant functional genomics and proteomics. *The Plant Journal*, **45**(4), 616–629.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., *et al.* (2004). Least angle regression. *The Annals of statistics*, **32**(2), 407–499.

Ellis, C. M., Nagpal, P., Young, J. C., Hagen, G., Guilfoyle, T. J., and Reed, J. W. (2005). AUXIN RESPONSE FACTOR1 and AUXIN RESPONSE FACTOR2 regulate senescence and floral organ abscission in Arabidopsis thaliana. *Development*, **132**(20), 4563–4574.

Elowitz, M. B. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**(6767), 335.

Engler, C., Kandzia, R., and Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. *PLoS One*, **3**(11), e3647.

Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009). Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One*, **4**(5), e5553.

Epple, P., Mack, A. A., Morris, V. R., and Dangl, J. L. (2003). Antagonistic control of oxidative stress-induced cell death in Arabidopsis by two related, plant-specific zinc finger proteins. *Proceedings of the National Academy of Sciences*, **100**(11), 6831–6836.

Eulgem, T. (2005). Regulation of the Arabidopsis defense transcriptome. *Trends in Plant Science*, **10**(2), 71–78.

Eulgem, T. and Somssich, I. E. (2007). Networks of WRKY transcription factors in defense signaling. *Current Opinion in Plant Biology*, **10**(4), 366–371.

Eulgem, T., Rushton, P. J., Robatzek, S., and Somssich, I. E. (2000). The WRKY superfamily of plant transcription factors. *Trends in Plant Science*, **5**(5), 199–206.

Evenson, R. E. and Gollin, D. (2003). Assessing the impact of the Green Revolution, 1960 to 2000. *science*, **300**(5620), 758–762.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, **5**(1), e8.

FAO (2011). Energy-smart food for people and climate (Issue Paper). Rome: Food and Agriculture Organization.

FAO and WFP (2012). The state of food insecurity in the world. page 65.

Faulkner, C., Petutschnig, E., Benitez-Alfonso, Y., Beck, M., Robatzek, S., Lipka, V., and Maule, A. J. (2013). LYM2-dependent chitin perception limits molecular flux via plasmodesmata. *Proceedings of the National Academy of Sciences*, **110**(22), 9166–9170.

Felix, G., Duran, J. D., Volko, S., and Boller, T. (1999). Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *The Plant Journal*, **18**(3), 265–276.

Ferrari, S., Plotnikova, J. M., De Lorenzo, G., and Ausubel, F. M. (2003). Arabidopsis local resistance to Botrytis cinerea involves salicylic acid and camalexin and requires EDS4 and PAD2, but not SID2, EDS5 or PAD4. *The Plant Journal*, **35**(2), 193–205.

Ferrari, S., Galletti, R., Denoux, C., De Lorenzo, G., Ausubel, F. M., and Dewdney, J. (2007). Resistance to Botrytis cinerea induced in Arabidopsis by elicitors is independent of salicylic acid, ethylene, or jasmonate signaling but requires PHYTOALEXIN DEFICIENT3. *Plant Physiology*, **144**(1), 367–379.

Ferrier, T., Matus, J. T., Jin, J., and Riechmann, J. L. (2011). Arabidopsis paves the way: genomic and network analyses in crops. *Current Opinion inBiotechnology*, **22**(2), 260–270.

Fethe, M. H., Liu, W., Burris, J. N., Millwood, R. J., Mazarei, M., Rudis, M. R., Yeaman, D. G., Dubosquielle, M., and Stewart, C. N. (2014). The performance of pathogenic bacterial phytosensing transgenic tobacco in the field. *Plant Biotechnology Journal*, **12**(6), 755–764.

Finkers, R., van den Berg, P., van Berloo, R., Ten Have, A., van Heusden, A. W., van Kan, J. A., and Lindhout, P. (2007). Three QTLs for Botrytis cinerea resistance in tomato. *Theoretical and Applied Genetics*, **114**(4), 585–593.

Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L., and Gurr, S. J. (2012). Emerging fungal threats to animal, plant and ecosystem health. *Nature*, **484**(7393), 186.

Fogelmark, K. and Troein, C. (2014). Rethinking transcriptional activation in the Arabidopsis circadian clock. *PLoS Computational Biology*, **10**(7), e1003705.

Foo, M., Gherman, I., Denby, K. J., and Bates, D. G. (2017). Control strategies for mitigating the effect of external perturbations on gene regulatory networks. *IFAC-PapersOnLine*, **50**(1), 12647–12652.

Franco-Zorrilla, J. M. and Solano, R. (2017). Identification of plant transcription factor target sequences. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1860**(1), 21–30.

Franco-Zorrilla, J. M., López-Vidriero, I., Carrasco, J. L., Godoy, M., Vera, P., and Solano, R. (2014). DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences*, **111**(6), 2367–2372.

Franklin, G., Powell, J., and Emami-Naeini, A. (2015). *Feedback control of dynamic systems*. Pearson, 7th edition.

Friso, G., Giacomelli, L., Ytterberg, A. J., Peltier, J.-B., Rudella, A., Sun, Q., and van Wijk, K. J. (2004). In-depth analysis of the thylakoid membrane proteome of Arabidopsis thaliana chloroplasts: new proteins, new functions, and a plastid proteome database. *The Plant Cell*, **16**(2), 478–499.

Fujita, M., Fujita, Y., Noutoshi, Y., Takahashi, F., Narusaka, Y., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2006). Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. *Current Opinion in Plant Biology*, **9**(4), 436–442.

Galon, Y., Nave, R., Boyce, J. M., Nachmias, D., Knight, M. R., and Fromm, H. (2008). Calmodulin-binding transcription activator (CAMTA) 3 mediates biotic defense responses in Arabidopsis. *FEBS letters*, **582**(6), 943–948.

García, A. V., Blanvillain-Baufumé, S., Huibers, R. P., Wiermer, M., Li, G., Gobbato, E., Rietz, S., and Parker, J. E. (2010). Balanced nuclear and cytoplasmic activities of EDS1 are required for a complete plant innate immune response. *PLoS pathogens*, **6**(7), e1000970.

Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in Escherichia coli. *Nature*, **403**(6767), 339.

Gardner, T. S., di Bernardo, D., Lorenz, D., and Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**(5629), 102–105.

Gassmann, W. and Bhattacharjee, S. (2012). Effector-triggered immunity signaling: from gene-for-gene pathways to protein-protein interaction networks. *Molecular Plant-Microbe Interactions*, **25**(7), 862–868.

Ge, L., Zhang, H., Chen, K., Ma, L., and Xu, Z. (2010). Effect of chitin on the antagonistic activity of Rhodotorula glutinis against Botrytis cinerea in strawberries and the possible mechanisms involved. *Food Chemistry*, **120**(2), 490–495.

Geddes, B. A., Ryu, M.-H., Mus, F., Costas, A. G., Peters, J. W., Voigt, C. A., and Poole, P. (2015). Use of plant colonizing bacteria as chassis for transfer of N 2-fixation to cereals. *Current Opinion inbiotechnology*, **32**, 216–222.

Georgianna, D. R. and Mayfield, S. P. (2012). Exploiting diversity and synthetic biology for the production of algal biofuels. *Nature*, **488**(7411), 329–335.

Ghassemian, M., Nambara, E., Cutler, S., Kawaide, H., Kamiya, Y., and McCourt, P. (2000). Regulation of abscisic acid signaling by the ethylene response pathway in Arabidopsis. *The Plant Cell*, **12**(7), 1117–1126.

Gifford, M. L., Dean, A., Gutierrez, R. A., Coruzzi, G. M., and Birnbaum, K. D. (2008). Cell-specific nitrogen responses mediate developmental plasticity. *Proceedings of the National Academy of Sciences*, **105**(2), 803–808.

Gigolashvili, T., Berger, B., Mock, H.-P., Müller, C., Weisshaar, B., and Flügge, U.-I. (2007). The transcription factor HIG1/MYB51 regulates indolic glucosinolate biosynthesis in Arabidopsis thaliana. *The Plant Journal*, **50**(5), 886–901.

Glazebrook, J. (2005). Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annual Review of Phytopathology*, **43**, 205–227.

Glenn, W. S., Runguphan, W., and O'Connor, S. E. (2013). Recent progress in the metabolic engineering of alkaloids in plant systems. *Current Opinion inbiotechnology*, **24**(2), 354–365.

Glickmann, E., Gardan, L., Jacquet, S., Hussain, S., Elasri, M., Petit, A., and Dessaux, Y. (1998). Auxin production is a common feature of most pathovars of Pseudomonas syringae. *Molecular Plant-Microbe Interactions*, **11**(2), 156–162.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., and Toulmin, C. (2010). Food security: the challenge of feeding 9 billion people. *science*, **327**(5967), 812–818.

Godoy, M., Franco-Zorrilla, J. M., Pérez-Pérez, J., Oliveros, J. C., Lorenzo, Ó., and Solano, R. (2011). Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. *The plant journal*, **66**(4), 700–711.

Gómez-Gómez, L. and Boller, T. (2000). FLS2: an LRR receptor–like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. *Molecular Cell*, **5**(6), 1003–1011.

Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A., and Carroll, S. B. (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*, **433**(7025), 481–487.

Gong, W., He, K., Covington, M., Dinesh-Kumar, S., Snyder, M., Harmer, S. L., Zhu, Y.-X., and Deng, X. W. (2008). The development of protein microarrays and their applications in DNA–protein and protein–protein interaction analyses of Arabidopsis transcription factors. *Molecular Plant*, **1**(1), 27–41.

Gordon, L. J., Finlayson, C. M., and Falkenmark, M. (2010). Managing water in agriculture for food production and other ecosystem services. *Agricultural Water Management*, **97**(4), 512–519.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, **4**(9), 117.

Greenfield, A., Hafemeister, C., and Bonneau, R. (2013). Robust data-driven incorpo-

ration of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, **29**(8), 1060–1067.

Großkopf, T. and Soyer, O. S. (2014). Synthetic microbial communities. *Current Opinion in Microbiology*, **18**, 72–77.

Guédon, Y., Barthélémy, D., Caraglio, Y., and Costes, E. (2001). Pattern analysis in branching and axillary flowering sequences. *Journal of Theoretical Biology*, **212**(4), 481–520.

Guelzim, N., Bottani, S., Bourgine, P., and Képès, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, **31**(1), 60.

Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L., and Luo, J. (2005). DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, **21**(10), 2568–2569.

Guo, F., Hanneke, S., Fu, W., and Xing, E. P. (2007). Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th international conference on Machine learning*, pages 321–328. ACM.

Guo, H. and Ecker, J. R. (2004). The ethylene signaling pathway: new insights. *Current Opinion in Plant Biology*, **7**(1), 40–49.

Ham, T. S., Dmytriv, Z., Plahar, H., Chen, J., Hillson, N. J., and Keasling, J. D. (2012). Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. *Nucleic Acids Research*, **40**(18), e141–e141.

Hamilton, A. J. and Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, **286**(5441), 950–952.

Hansen, A. S. and O'shea, E. K. (2013). Promoter decoding of transcription factor dynamics involves a trade-off between noise and control of gene expression. *Molecular Systems Biology*, **9**(1), 704.

Harris, A. W. K., Dolan, J. A., Kelly, C. L., Anderson, J., and Papachristodoulou, A. (2015). Designing genetic feedback controllers. *IEEE Transactions on Biomedical Circuits and Systems*, **9**(4), 475–484.

Hasson, A., Plessis, A., Blein, T., Adroher, B., Grigg, S., Tsiantis, M., Boudaoud, A., Damerval, C., and Laufs, P. (2011). Evolution and diverse roles of the CUP-SHAPED COTYLEDON genes in Arabidopsis leaf development. *The Plant Cell*, page 110.

Hauck, P., Thilmony, R., and He, S. Y. (2003). A Pseudomonas syringae type III effector suppresses cell wall-based extracellular defense in susceptible Arabidopsis plants. *Proceedings of the National Academy of Sciences*, **100**(14), 8577–8582.

Hauptmann, V., Weichert, N., Rakhimova, M., and Conrad, U. (2013). Spider silks from plants–a challenge to create native-sized spidroins. *Biotechnology Journal*, **8**(10), 1183–1192.

Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). TIGRESS: trustful inference of gene regulation using stability selection. *BMC Systems Biology*, **6**(145).

Häusler, R. E., Hirsch, H.-J., Kreuzaler, F., and Peterhänsel, C. (2002). Overexpression of C4-cycle enzymes in transgenic C3 plants: a biotechnological approach to improve C3-photosynthesis. *Journal of Experimental Botany*, **53**(369), 591–607.

Hayat, S., Hayat, Q., Alyemeni, M. N., Wani, A. S., Pichtel, J., and Ahmad, A. (2012).

Role of proline under changing environments: a review. *Plant Signaling & Behavior*, **7**(11), 1456–1466.

He, H. and Siu, W.-C. (2011). Single image super-resolution using Gaussian process regression. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 449–456. IEEE.

He, X. and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genetics*, **2**(6), e88.

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models - a review. *Biosystems*, **96**(1), 86–103.

Heidel, A. J., Clarke, J. D., Antonovics, J., and Dong, X. (2004). Fitness costs of mutations affecting the systemic acquired resistance pathway in Arabidopsis thaliana. *Genetics*, **168**(4), 2197–2206.

Heiner, M., Gilbert, D., and Donaldson, R. (2008). Petri nets for systems and synthetic biology. In *International school on formal methods for the design of computer, communication and software systems*, pages 215–264. Springer.

Herrero, E., Kolmos, E., Bujdoso, N., Yuan, Y., Wang, M., Berns, M. C., Uhlworm, H., Coupland, G., Saini, R., Jaskolski, M., Webb, A., Gonçalves, J., and Davis, S. J. (2012). EARLY FLOWERING4 recruitment of EARLY FLOWERING3 in the nucleus sustains the Arabidopsis circadian clock. *The Plant Cell*, **24**(2), 428–443.

Hespanha, J. P., Naghshtabrizi, P., and Xu, Y. (2007). A survey of recent results in networked control systems. *Proceedings of the IEEE*, **95**(1), 138–162.

Heyndrickx, K. S., Van de Velde, J., Wang, C., Weigel, D., and Vandepoele, K. (2014). A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana. *The Plant Cell*, **26**(10), 3894–3910.

Heyne, E., Brunson, A. M., *et al.* (1940). Genetic studies of heat and drought tolerance in maize. *Journal of the American Society of Agronomy*, **32**, 803–14.

Hickman, R., Hill, C., Penfold, C. A., Breeze, E., Bowden, L., Moore, J. D., Zhang, P., Jackson, A., Cooke, E., Bewicke-Copley, F., Mead, A., Beynon, J., Wild, D. L., Denby, K. J., Ott, S., and Buchanan-Wollaston, V. (2013). A local regulatory network around three NAC transcription factors in stress responses and senescence in Arabidopsis leaves. *The Plant Journal*, **75**(1), 26–39.

Hickman, R., van Verk, M. C., Van Dijken, A. J., Pereira Mendes, M., Vroegop-Vos, I. A., Caarls, L., Steenbergen, M., Van Der Nagel, I., Wesselink, G. J., Jironkin, A., Talbot, A., Rhodes, J., de Vries, M., Schuurink, R. C., Denby, K., Pieterse, C. M., and Van Wees, S. C. (2017). Architecture and dynamics of the jasmonic acid gene regulatory network. *The Plant Cell*, pages Aug, 2017. DOI:10.1105/tpc.16.00958.

Higashi, K., Ishiga, Y., Inagaki, Y., Toyoda, K., Shiraishi, T., and Ichinose, Y. (2008). Modulation of defense signal transduction by flagellin-induced WRKY41 transcription factor in Arabidopsis thaliana. *Molecular Genetics and Genomics*, **279**(3), 303–312.

Hill, K., Porco, S., Lobet, G., Zappala, S., Mooney, S., Draye, X., and Bennett, M. J. (2013). Root systems biology: integrative modeling across scales, from gene regulatory networks to the rhizosphere. *Plant Physiology*, **163**(4), 1487–1503.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Hofmann, N. R. (2013). A NAC Transcription Factor for Flooding: SHYG Helps Plants Keep Their Leaves in the Air.

Holt III, B. F., Boyes, D. C., Ellerström, M., Siefers, N., Wiig, A., Kauffman, S., Grant, M. R., and Dangl, J. L. (2002). An evolutionarily conserved mediator of plant disease resistance gene function is required for normal Arabidopsis development. *Developmental Cell*, **2**(6), 807–817.

Hong-bo, S., Li-ye, C., Chang-xing, Z., Qing-jie, G., Xian-an, L., and Jean-Marcel, R. (2006). Plant gene regulatory network system under abiotic stress. *Acta Biologica Szegediensis*, **50**(1-2), 1–9.

Hornitschek, P., Kohnen, M. V., Lorrain, S., Rougemont, J., Ljung, K., López-Vidriero, I., Franco-Zorrilla, J. M., Solano, R., Trevisan, M., Pradervand, S., *et al.* (2012). Phytochrome interacting factors 4 and 5 control seedling growth in changing light conditions by directly controlling auxin signaling. *The Plant Journal*, **71**(5), 699–711.

Hsieh, T.-H., Lee, J.-T., Yang, P.-T., Chiu, L.-H., Charng, Y.-y., Wang, Y.-C., and Chan, M.-T. (2002). Heterology expression of the ArabidopsisC-repeat/dehydration response element binding Factor 1 gene confers elevated tolerance to chilling and oxidative stresses in transgenic tomato. *Plant Physiology*, **129**(3), 1086–1094.

Hsu, C.-Y., Adams, J. P., Kim, H., No, K., Ma, C., Strauss, S. H., Drnevich, J., Vandervelde, L., Ellis, J. D., Rice, B. M., *et al.* (2011). FLOWERING LOCUS T duplication coordinates reproductive and vegetative growth in perennial poplar. *Proceedings of the National Academy of Sciences*, **108**(26), 10756–10761.

Hua, J., Chang, C., Sun, Q., and Meyerowitz, E. M. (1995). Ethylene insensitivity conferred by Arabidopsis ERS gene. *Science*, **269**(5231), 1712–1714.

Huang, X. N. and Ren, H. P. (2017). Understanding robust adaptation dynamics of gene regulatory network. *IEEE Transactions on Biomedical Circuits and Systems*, **11**(4), 942–957.

Hückelhoven, R. (2007). Cell wall–associated mechanisms of disease resistance and susceptibility. *Annual Review of Phytopathology*, **45**, 101–127.

Hughes, M. E., Hogenesch, J. B., and Kornacker, K. (2010). JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of biological rhythms*, **25**(5), 372–380.

Hunt, R. W. and Wasserman, W. W. (2014). Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biology*, **15**(7), 412.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**(9), e12776.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, **8**(1), 565.

Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J., and Barkai, N. (2005). Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science*, **309**(5736), 938–940.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A., and Shinozaki, K. (2005). RARTF: database and tools for complete sets of Arabidopsis transcription factors. *DNA Research*, **12**(4), 247–256.

International, R. G. S. P. (2005). The map-based sequence of the rice genome. *Nature*, **436**(7052), 793.

Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., Garriga-Canut, M., and Serrano, L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, **452**(7189), 840–845.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & Cellular Proteomics*, **4**(9), 1265–1272.

Iwase, A., Matsui, K., and Ohme-Takagi, M. (2009). Manipulation of plant metabolic pathways by transcription factors. *Plant Biotechnology*, **26**(1), 29–38.

Jamir, Y., Guo, M., Oh, H.-S., Petnicki-Ocwieja, T., Chen, S., Tang, X., Dickman, M. B., Collmer, A., and R. Alfano, J. (2004). Identification of Pseudomonas syringae type III effectors that can suppress programmed cell death in plants and yeast. *The Plant Journal*, **37**(4), 554–565.

Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**(6833), 41.

Jiang, C.-Z. (2004). Method for modifying plant biomass. US Patent 6,717,034.

Jiang, H., Li, H., Bu, Q., and Li, C. (2009). The RHA2a-interacting proteins ANAC019 and ANAC055 may play a dual role in regulating ABA response and jasmonate response. *Plant Signaling & Behavior*, **4**(5), 464–466.

Jiang, Y. and Huang, B. (2001). Drought and Heat Stress Injury to Two Cool-Season Turfgrasses in Relation to Antioxidant Metabolism and Lipid Peroxidation Contribution No. 00-227-J from Kansas Agric. Exp. Stn. *Crop Science*, **41**(2), 436–442.

Jin, J., He, K., Tang, X., Li, Z., Lv, L., Zhao, Y., Luo, J., and Gao, G. (2015). An Arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors. *Molecular Biology and Evolution*, **32**(7), 1767–1773.

Jin, W., Wu, F., Xiao, L., Liang, G., Zhen, Y., Guo, Z., and Guo, A. (2012). Microarray-based analysis of tomato miRNA regulated by Botrytis cinerea. *Journal of Plant Growth Regulation*, **31**(1), 38–46.

Jones, J. D. G. and Dangl, J. L. (2006). The plant immune system. *Nature*, **444**(7117), 323–329.

Ju, C., Yoon, G. M., Shemansky, J. M., Lin, D. Y., Ying, Z. I., Chang, J., Garrett, W. M., Kessenbrock, M., Groth, G., Tucker, M. L., *et al.* (2012). CTR1 phosphorylates the central regulator EIN2 to control ethylene hormone signaling from the ER membrane to the nucleus in Arabidopsis. *Proceedings of the National Academy of Sciences*, **109**(47), 19486–19491.

Kaminaka, H., Näke, C., Epple, P., Dittgen, J., Schütze, K., Chaban, C., Holt, B. F., Merkle, T., Schäfer, E., Harter, K., *et al.* (2006). bZIP10-LSD1 antagonism modulates

basal defense and cell death in Arabidopsis following infection. *The EMBO journal*, **25**(18), 4400–4411.

Kanamori, M., Konno, H., Osato, N., Kawai, J., Hayashizaki, Y., and Suzuki, H. (2004). A genome-wide and nonredundant mouse transcription factor database. *Biochemical and Biophysical Research Communications*, **322**(3), 787–793.

Kang, H., Park, S. J., and Kwak, K. J. (2013). Plant RNA chaperones in stress response. *Trends in Plant Science*, **18**(2), 100–106.

Karim, S., Aronsson, H., Ericson, H., Pirhonen, M., Leyman, B., Welin, B., Mäntylä, E., Palva, E. T., Van Dijck, P., and Holmström, K.-O. (2007). Improved drought tolerance without undesired side effects in transgenic plants producing trehalose. *Plant Molecular Biology*, **64**(4), 371–386.

Karimi, M., Inzé, D., and Depicker, A. (2002). GATEWAY vectors for Agrobacterium-mediated plant transformation. *Trends in Plant Science*, **7**(5), 193–195.

Kasuga, M., Liu, Q., Miura, S., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1999). Improving plant drought, salt, and freezing tolerance by gene transfer of a single stress-inducible transcription factor. *Nature Biotechnology*, **17**(3), 287.

Katagiri, F. (2004). A global view of defense gene expression regulation–a highly interconnected signaling network. *Current Opinion in Plant Biology*, **7**(5), 506–511.

Kaufmann, K., Muino, J. M., Østerås, M., Farinelli, L., Krajewski, P., and Angenent, G. C. (2010). Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nature Protocols*, **5**(3), 457–472.

Kay, R., Chan, A., Daly, M., and McPherson, J. (1987). Duplication of CaMV 35S promoter sequences creates a strong enhancer for plant genes. *Science*, **236**(4806), 1299–1302.

Kazan, K. and Manners, J. M. (2012). JAZ repressors and the orchestration of phytohormone crosstalk. *Trends in Plant Science*, **17**(1), 22–31.

Kazan, K. and Manners, J. M. (2013). MYC2: the master in action. *Molecular Plant*, **6**(3), 686–703.

Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñiz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., *et al.* (2010). EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research*, **39**(suppl_1), D583–D590.

Kevei, E., Gyula, P., Hall, A., Kozma-Bognár, L., Kim, W.-Y., Eriksson, M. E., Tóth, R., Hanano, S., Fehér, B., Southern, M. M., *et al.* (2006). Forward genetic analysis of the circadian clock separates the multiple functions of ZEITLUPE. *Plant Physiology*, **140**(3), 933–945.

Khanin, R. and Wit, E. (2006). How scale-free are biological networks. *Journal of Computational Biology*, **13**(3), 810–818.

Khanna, R., Kronmiller, B., Maszle, D. R., Coupland, G., Holm, M., Mizuno, T., and Wu, S.-H. (2009). The Arabidopsis B-box zinc finger family. *The Plant Cell*, **21**(11), 3416–3420.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'angelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K. (2007). The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *The Plant Journal*, **50**(2), 347–363.

Kim, J., Bates, D. G., Postlethwaite, I., Heslop-Harrison, P., and Cho, K.-H. (2007). Least-squares methods for identifying biochemical regulatory networks from noisy measurements. *BMC Bioinformatics*, **8**(1), 8.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H. (2008a). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**(6), 1049–1061.

Kim, J., Bates, D. G., Postlethwaite, I., Heslop-Harrison, P., and Cho, K.-H. (2008b). Linear time-varying models can reveal non-linear interactions of biomolecular regulatory networks using multiple time-series data. *Bioinformatics*, **24**(10), 1286–1292.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, **5**(11), 826.

Klemm, S. *et al.* (2008). Causal structure identification in nonlinear dynamical systems. *Department of Engineering, University of Cambridge, UK*.

Koike, N., Yoo, S.-H., Huang, H.-C., Kumar, V., Lee, C., Kim, T.-K., and Takahashi, J. S. (2012). Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science*, **338**(6105), 349–354.

Koiwa, H., Bressan, R. A., and Hasegawa, P. M. (2006). Identification of plant stress-responsive determinants in Arabidopsis by large-scale forward genetic screens. *Journal of Experimental Botany*, **57**(5), 1119–1128.

Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D., and Church, G. M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences*, **110**(34), 14024–14029.

Kreps, J. A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X., and Harper, J. F. (2002). Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiology*, **130**(4), 2129–2141.

Krishnakumar, V., Hanlon, M. R., Contrino, S., Ferlanti, E. S., Karamycheva, S., Kim, M., Rosen, B. D., Cheng, C.-Y., Moreira, W., Mock, S. A., Stubbs, J., Sullivan, J. M., Krampis, K., Miller, J. R., Micklem, G., Vaughn, M., and Town, C. D. (2015). Araport: the Arabidopsis information portal. *Nucleic Acids Research*, **43**(D1), D1003–D1009.

Krishnakumar, V., Contrino, S., Cheng, C.-Y., Belyaeva, I., Ferlanti, E. S., Miller, J. R., Vaughn, M. W., Micklem, G., Town, C. D., and Chan, A. P. (2016). ThaleMine: a warehouse for Arabidopsis data integration and discovery. *Plant and Cell Physiology*, **58**(1), e4–e4.

Krishnaswamy, S., Verma, S., Rahman, M. H., and Kav, N. N. (2011). Functional characterization of four APETALA2-family genes (RAP2. 6, RAP2. 6L, DREB19 and DREB26) in Arabidopsis. *Plant Molecular Biology*, **75**(1-2), 107–127.

Krizek, B. A., Bequette, C. J., Xu, K., Blakley, I. C., Fu, Z. Q., Stratmann, J., and Loraine, A. E. (2016). RNA-Seq links AINTEGUMENTA and AINTEGUMENTA-LIKE6 to cell wall remodeling and plant defense pathways in Arabidopsis. *Plant Physiology*, pages pp–01625.

Krysan, P. J., Young, J. C., and Sussman, M. R. (1999). T-DNA as an insertional mutagen in Arabidopsis. *The Plant Cell*, **11**(12), 2283–2290.

Kulkarni, M. M. (2011). Digital multiplexed gene expression analysis using the NanoString nCounter system. *Current Protocols in Molecular Biology*, **94**(1), 25B–10.

Kulkarni, S. R., Vaneechoutte, D., Van de Velde, J., and Vandepoele, K. (2017). TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Research*, **46**(6), e31–e31.

Kurt, Bogaert, T., Coddens, K., Deschouwer, K., Van Hummelen, P., Vuylsteke, M., Moreau, Y., Kwekkeboom, J., Wijfjes, A. H., May, S., Beynon, J., Hilson, P., and Kuiper, M. T. (2005). Benchmarking the CATMA microarray. A novel tool for Arabidopsis transcriptome analysis. *Plant Physiology*, **137**(2), 588–601.

Lai, H., Engle, M., Fuchs, A., Keller, T., Johnson, S., Gorlatov, S., Diamond, M. S., and Chen, Q. (2010). Monoclonal antibody produced in plants efficiently treats West Nile virus infection in mice. *Proceedings of the National Academy of Sciences*, **107**(6), 2419–2424.

Lai, Z., Vinod, K., Zheng, Z., Fan, B., and Chen, Z. (2008). Roles of Arabidopsis WRKY3 and WRKY4 transcription factors in plant responses to pathogens. *BMC Plant Biology*, **8**(1), 68.

Laluk, K., AbuQamar, S., and Mengiste, T. (2011). The Arabidopsis mitochondrial localized pentatricopeptide repeat protein PGN functions in defense against necrotrophic fungi and abiotic stress tolerance. *Plant Physiology*, pages pp–111.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., *et al.* (2011). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40**(D1), D1202–D1210.

Lan, P., Li, W., and Schmidt, W. (2012). Complementary proteome and transcriptome profiling in phosphate-deficient Arabidopsis roots reveals multiple levels of gene regulation. *Molecular & Cellular Proteomics*, **11**(11), 1156–1166.

Lata, C. and Prasad, M. (2011). Role of DREBs in regulation of abiotic stress responses in plants. *Journal of Experimental Botany*, **62**(14), 4731–4748.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, **15**(2), R29.

Lawrence, N. D., Sanguinetti, G., and Rattray, M. (2007). Modelling transcriptional regulation using Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 785–792.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., and Rhee, S. Y. (2010). Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nature Biotechnology*, **28**(2), 149.

Lee, S. C. and Luan, S. (2012). ABA signal transduction at the crossroad of biotic and abiotic stress responses. *Plant, Cell & Environment*, **35**(1), 53–60.

Lelièvre, J.-M., Latchè, A., Jones, B., Bouzayen, M., and Pech, J.-C. (1997). Ethylene and fruit ripening. *Physiologia Plantarum*, **101**(4), 727–739.

LeNoble, M. E., Spollen, W. G., and Sharp, R. E. (2004). Maintenance of shoot growth by endogenous ABA: genetic assessment of the involvement of ethylene suppression. *Journal of Experimental Botany*, **55**(395), 237–245.

Leroux, P., Fritz, R., Debieu, D., Albertini, C., Lanen, C., Bach, J., Gredt, M., and Chapeland, F. (2002). Mechanisms of resistance to fungicides in field strains of Botrytis cinerea. *Pest Management Science*, **58**(9), 876–888.

Lewis, L. A., Polanski, K., de Torres-Zabala, M., Jayaraman, S., Bowden, L., Moore, J., Penfold, C. A., Jenkins, D. J., Hill, C., Baxter, L., Kulasekaran, S., Truman, W., Littlejohn, G., Prusinska, J., Mead, A., Steinbrenner, J., Hickman, R., Rand, D., Wild, D. L., Ott, S., Buchanan-Wollaston, V., Smirnoff, N., Beynon, J., Denby, K., and Grant, M. (2015). Transcriptional dynamics driving MAMP-triggered immunity and pathogen effector-mediated immunosuppression in Arabidopsis leaves following infection with Pseudomonas syringae pv tomato DC3000. *The Plant Cell*, **27**(11), 3038–3064.

Leyva-González, M. A., Ibarra-Laclette, E., Cruz-Ramírez, A., and Herrera-Estrella, L. (2012). Functional and transcriptome analysis reveals an acclimatization strategy for abiotic stress tolerance mediated by Arabidopsis NF-YA family members. *PLoS One*, **7**(10), e48138.

Li, J. J., Bickel, P. J., and Biggin, M. D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, **2**, e270.

Li, W.-X., Oono, Y., Zhu, J., He, X.-J., Wu, J.-M., Iida, K., Lu, X.-Y., Cui, X., Jin, H., and Zhu, J.-K. (2008). The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. *The Plant Cell*, **20**(8), 2238–2251.

Li, Y. and Patra, J. C. (2010). Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**(9), 1219–1224.

Li, Y.-J., Fang, Y., Fu, Y.-R., Huang, J.-G., Wu, C.-A., and Zheng, C.-C. (2013). NFYA1 is involved in regulation of postgermination growth arrest under salt stress in Arabidopsis. *PLoS One*, **8**(4), e61289.

Libault, M., Wan, J., Czechowski, T., Udvardi, M., and Stacey, G. (2007). Identification of 118 Arabidopsis transcription factor and 30 ubiquitin-ligase genes responding to chitin, a plant-defense elicitor. *Molecular Plant-Microbe Interactions*, **20**(8), 900–911.

Libault, M., Brechenmacher, L., Cheng, J., Xu, D., and Stacey, G. (2010). Root hair systems biology. *Trends in Plant Science*, **15**(11), 641–650.

Libault, M., Pingault, L., Zogli, P., and Schiefelbein, J. (2017). Plant systems biology at the single-cell level. *Trends in Plant Science*.

Lienert, F., Lohmueller, J. J., Garg, A., and Silver, P. A. (2014). Synthetic Biology in mammalian cells: next generation research tools and therapeutics. *Nature reviews. Molecular Cell Biology*, **15**(2), 95.

Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C. P., and Stumpf, M. P. (2014). A

framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nature Protocols*, **9**(2), 439.

Lillacci, G., Aoki, S. K., Schweingruber, D., and Khammash, M. (2017). A synthetic integral feedback controller for robust tunable regulation in bacteria. *bioRxiv*, page 170951.

Lima-Mendez, G. and van Helden, J. (2009). The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, **5**(12), 1482–1493.

Lin, Z., Griffith, M. E., Li, X., Zhu, Z., Tan, L., Fu, Y., Zhang, W., Wang, X., Xie, D., and Sun, C. (2007). Origin of seed shattering in rice (Oryza sativa L.). *Planta*, **226**(1), 11–20.

Liu, F., Jiang, H., Ye, S., Chen, W.-P., Liang, W., Xu, Y., Sun, B., Sun, J., Wang, Q., Cohen, J. D., *et al.* (2010). The Arabidopsis P450 protein CYP82C2 modulates jasmonate-induced root growth inhibition, defense gene expression and indole glucosinolate biosynthesis. *Cell Research*, **20**(5), 539.

Liu, J.-X. and Howell, S. H. (2010). bZIP28 and NF-Y transcription factors are activated by ER stress and assemble into a transcriptional complex to regulate stress response genes in Arabidopsis. *The Plant Cell*, **22**(3), 782–796.

Liu, Q., Kasuga, M., Sakuma, Y., Abe, H., Miura, S., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1998). Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought-and low-temperature-responsive gene expression, respectively, in Arabidopsis. *The Plant Cell*, **10**(8), 1391–1406.

Liu, S., Kracher, B., Ziegler, J., Birkenbihl, R. P., and Somssich, I. E. (2015). Negative regulation of ABA signaling by WRKY33 is critical for Arabidopsis immunity towards Botrytis cinerea 2100. *Elife*, **4**, e07295.

Liu, W. and Stewart, C. N. (2015). Plant synthetic biology. *Trends in Plant Science*, **20**(5), 309–317.

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. (2011). Controllability of complex networks. *Nature*, **473**(7346), 167–173.

Liu, Z., Malone, B., and Yuan, C. (2012). Empirical evaluation of scoring functions for Bayesian network model selection. In *BMC Bioinformatics*, volume 13, page S14. BioMed Central.

Lizotte, D. J., Wang, T., Bowling, M. H., and Schuurmans, D. (2007). Automatic Gait Optimization with Gaussian Process Regression. In *IJCAI*, volume 7, pages 944–949.

Lloyd, J. R., Duvenaud, D. K., Grosse, R. B., Tenenbaum, J. B., and Ghahramani, Z. (2014). Automatic Construction and Natural-Language Description of Nonparametric Regression Models. In *AAAI*, pages 1242–1250.

Lo Presti, L., Lanver, D., Schweizer, G., Tanaka, S., Liang, L., Tollot, M., Zuccaro, A., Reissmann, S., and Kahmann, R. (2015). Fungal effectors and plant susceptibility. *Annual review of Plant Biology*, **66**, 513–545.

Loake, G. and Grant, M. (2007). Salicylic acid in plant defence - the players and protagonists. *Current Opinion in Plant Biology*, **10**(5), 466–472.

Locke, J. C., Kozma-Bognár, L., Gould, P. D., Fehér, B., Kevei, E., Nagy, F., Turner, M. S., Hall, A., and Millar, A. J. (2006). Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. *Molecular Systems Biology*, **2**(1), 59.

Long, S. P., Marshall-Colon, A., and Zhu, X.-G. (2015). Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. *Cell*, **161**(1), 56–66.

Loraine, A. E., McCormick, S., Estrada, A., Patel, K., and Qin, P. (2013). RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. *Plant Physiology*, **162**(2), 1092–1109.

Lorenzo, O., Piqueras, R., Sánchez-Serrano, J. J., and Solano, R. (2003). ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense. *The Plant Cell*, **15**(1), 165–178.

Lorenzo, O., Chico, J. M., Sánchez-Serrano, J. J., and Solano, R. (2004). JASMONATE-INSENSITIVE1 encodes a MYC transcription factor essential to discriminate between different jasmonate-regulated defense responses in Arabidopsis. *The Plant Cell*, **16**(7), 1938–1950.

Lorković, Z. J. (2009). Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends in Plant Science*, **14**(4), 229–236.

Lucas, M., Laplaze, L., and Bennett, M. J. (2011). Plant systems biology: network matters. *Plant, Cell & Environment*, **34**(4), 535–553.

Luhua, S., Hegie, A., Suzuki, N., Shulaev, E., Luo, X., Cenariu, D., Ma, V., Kao, S., Lim, J., Gunay, M. B., *et al.* (2013). Linking genes of unknown function with abiotic stress responses by high-throughput phenotype screening. *Physiologia Plantarum*, **148**(3), 322–333.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**(7006), 308–312.

Ma, W., Trusina, A., El-Samad, H., Lim, W. A., and Tang, C. (2009). Defining network topologies that can achieve biochemical adaptation. *Cell*, **138**(4), 760–773.

Madar, A., Greenfield, A., Vanden-Eijnden, E., and Bonneau, R. (2010). DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS One*, **5**(3), e9803.

Madsen, C., McLaughlin, J. A., Mısırlı, G., Pocock, M., Flanagan, K., Hallinan, J., and Wipat, A. (2016). The SBOL Stack: a platform for storing, publishing, and sharing synthetic biology designs. *ACS Synthetic Biology*, **5**(6), 487–497.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS letters*, **583**(24), 3966–3973.

Marbach, D., Schaffter, T., Floreano, D., Prill, R. J., and Stolovitzky, G. (2009a). The DREAM4 in-silico network challenge. *Draft, version 0.3*.

Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. (2009b). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of computational Biology*, **16**(2), 229–239.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, **107**(14), 6286–6291.

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, **9**(8), 796–804.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC Bioinformatics*, volume 7, page S7. BioMed Central.

Markowetz, F. and Spang, R. (2007). Inferring cellular networks–a review. *BMC Bioinformatics*, **8**(6), 1.

Marquez, Y., Brown, J. W., Simpson, C. G., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research*, pages gr–134106.

Marti-Marimon, M., Vialaneix, N., Voillet, V., Yerle-Bouissou, M., Lahbib-Mansais, Y., and Liaubet, L. (2018). A new approach of gene co-expression network inference reveals significant biological processes involved in porcine muscle development in late gestation. *Scientific reports*, **8**(1), 10150.

Marx, V. (2016). Plants: a tool box of cell-based assays.

Mauch-Mani, B. and Mauch, F. (2005). The role of abscisic acid in plant–pathogen interactions. *Current Opinion in Plant Biology*, **8**(4), 409–414.

Mazid, M., Khan, T., Mohammad, F., *et al.* (2011). Role of secondary metabolites in defense mechanisms of plants. *Biology and Medicine*, **3**(2), 232–249.

Mazzucotelli, E., Mastrangelo, A. M., Crosatti, C., Guerra, D., Stanca, A. M., and Cattivelli, L. (2008). Abiotic stress response in plants: when post-transcriptional and post-translational regulations control transcription. *Plant Science*, **174**(4), 420–431.

McClung, C. R. (2006). Plant circadian rhythms. *The Plant Cell*, **18**(4), 792–803.

McGrath, K. C., Dombrecht, B., Manners, J. M., Schenk, P. M., Edgar, C. I., Maclean, D. J., Scheible, W.-R., Udvardi, M. K., and Kazan, K. (2005). Repressor-and activator-type ethylene response factors functioning in jasmonate signaling and disease resistance identified via a genome-wide screen of Arabidopsis transcription factor gene expression. *Plant Physiology*, **139**(2), 949–959.

McKenzie, F. C. and Williams, J. (2015). Sustainable food production: constraints, challenges and choices by 2050. *Food Security*, **7**(2), 221–233.

McLaughlin, J. A., Myers, C. J., Zundel, Z., Mısırlı, G., Zhang, M., Ofiteru, I. D., Goni-Moreno, A., and Wipat, A. (2018). SynBioHub: A Standards-Enabled Design Repository for Synthetic Biology. *ACS Synthetic Biology*, **7**(2), 682–688.

Mengiste, T., Chen, X., Salmeron, J., and Dietrich, R. (2003). The BOTRYTIS SUSCEPTIBLE1 gene encodes an R2R3MYB transcription factor protein that is required for biotic and abiotic stress responses in Arabidopsis. *The Plant Cell*, **15**(11), 2551–2565.

Menon, P. P., Postlethwaite, I., Bennani, S., Marcos, A., and Bates, D. G. (2009). Robustness analysis of a reusable launch vehicle flight control law. *Control Engineering Practice*, **17**(7), 751–765.

Meyer, R. S. and Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics*, **14**(12), 840.

Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, **8**(8), 1551.

Milias-Argeitis, A., Summers, S., Stewart-Ornstein, J., Zuleta, I., Pincus, D., El-Samad, H., Khammash, M., and Lygeros, J. (2011). In silico feedback for in vivo regulation of a gene expression circuit. *Nature Biotechnology*, **29**(12), 1114.

Millet, Y. A., Danna, C. H., Clay, N. K., Songnuan, W., Simon, M. D., Werck-Reichhart, D., and Ausubel, F. M. (2010). Innate immune responses activated in Arabidopsis roots by microbe-associated molecular patterns. *The Plant Cell*, **22**(3), 973–990.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**(5594), 824–827.

Mironova, V. V., Omelyanchuk, N. A., Yosiphon, G., Fadeev, S. I., Kolchanov, N. A., Mjolsness, E., and Likhoshvai, V. A. (2010). A plausible mechanism for auxin patterning along the developing root. *BMC Systems Biology*, **4**(1), 98.

Mitsuda, N. and Ohme-Takagi, M. (2009). Functional analysis of transcription factors in Arabidopsis. *Plant and Cell Physiology*, **50**(7), 1232–1248.

Mittler, R. (2006). Abiotic stress, the field environment and stress combination. *Trends in Plant Science*, **11**(1), 15–19.

Miya, A., Albert, P., Shinya, T., Desaki, Y., Ichimura, K., Shirasu, K., Narusaka, Y., Kawakami, N., Kaku, H., and Shibuya, N. (2007). CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis. *Proceedings of the National Academy of Sciences*, **104**(49), 19613–19618.

Mizoguchi, T., Wheatley, K., Hanzawa, Y., Wright, L., Mizoguchi, M., Song, H.-R., Carré, I. A., and Coupland, G. (2002). LHY and CCA1 are partially redundant genes required to maintain circadian rhythms in Arabidopsis. *Developmental Cell*, **2**(5), 629–641.

Moffat, C. S., Ingle, R. A., Wathugala, D. L., Saunders, N. J., Knight, H., and Knight, M. R. (2012). ERF5 and ERF6 play redundant roles as positive regulators of JA/Et-mediated defense against Botrytis cinerea in Arabidopsis. *PLoS One*, **7**(4), e35995.

Mohr, P. G. and Cahill, D. M. (2003). Abscisic acid influences the susceptibility of Arabidopsis thaliana to Pseudomonas syringae pv. tomato and Peronospora parasitica. *Functional Plant Biology*, **30**(4), 461–469.

Montes, R. A. C., Coello, G., González-Aguilera, K. L., Marsch-Martínez, N., de Folter, S., and Alvarez-Buylla, E. R. (2014). ARACNe-based inference, using curated microarray data, of Arabidopsis thaliana root transcriptional regulatory networks. *BMC Plant Biology*, **14**(1), 97.

Moore, R. C. and Purugganan, M. D. (2005). The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology*, **8**(2), 122–128.

Morrissey, E. R. (2013). GRENITS: Gene regulatory network inference using time series. *dim (Athaliana_ODE)*, **1**(5), 50.

Muhammad, D., Schmittling, S., Williams, C., and Long, T. A. (2017). More than meets the eye: Emergent properties of transcription factors networks in Arabidopsis. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1860**(1), 64–74.

Muraro, D., Mellor, N., Pound, M. P., Lucas, M., Chopard, J., Byrne, H. M., Godin, C., Hodgman, T. C., King, J. R., Pridmore, T. P., *et al.* (2014). Integration of hormonal signaling networks and mobile microRNAs is required for vascular patterning in Arabidopsis roots. *Proceedings of the National Academy of Sciences*, **111**(2), 857–862.

Murray-Smith, R., Johansen, T. A., and Shorten, R. (1999). On transient dynamics, off-equilibrium behaviour and identification in blended multiple model structures. In *Control Conference (ECC), 1999 European*, pages 3569–3574. IEEE.

Nägele, T., Henkel, S., Hörmiller, I., Sauter, T., Sawodny, O., Ederer, M., and Heyer, A. G. (2010). Mathematical modeling of the central carbohydrate metabolism in Arabidopsis reveals a substantial regulatory influence of vacuolar invertase on whole plant carbon metabolism. *Plant Physiology*, **153**(1), 260–272.

Naito, K., Taguchi, F., Suzuki, T., Inagaki, Y., Toyoda, K., Shiraishi, T., and Ichinose, Y. (2008). Amino acid sequence of bacterial microbe-associated molecular pattern flg22 is required for virulence. *Molecular Plant-Microbe Interactions*, **21**(9), 1165–1174.

Nakano, T., Suzuki, K., Fujimura, T., and Shinshi, H. (2006). Genome-wide analysis of the ERF gene family in Arabidopsis and rice. *Plant Physiology*, **140**(2), 411–432.

Narsai, R., Howell, K. A., Millar, A. H., O'Toole, N., Small, I., and Whelan, J. (2007). Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. *The Plant Cell*, **19**(11), 3418–3436.

Needham, C. J., Manfield, I. W., Bulpitt, A. J., Gilmartin, P. M., and Westhead, D. R. (2009). From gene expression to gene regulatory networks in Arabidopsis thaliana. *BMC Systems Biology*, **3**(1), 85.

Ner-Gaon, H., Leviatan, N., Rubin, E., and Fluhr, R. (2007). Comparative cross-species alternative splicing in plants. *Plant Physiology*, **144**(3), 1632–1641.

Newman, M.-A., Sundelin, T., Nielsen, J. T., and Erbs, G. (2013). MAMP (microbe-associated molecular pattern) triggered immunity in plants. *Frontiers in Plant Science*, **4**, 139.

Nguyen-Tuong, D., Seeger, M., and Peters, J. (2009). Model learning with local gaussian process regression. *Advanced Robotics*, **23**(15), 2015–2034.

Niu, Y., Qian, D., Liu, B., Ma, J., Wan, D., Wang, X., He, W., and Xiang, Y. (2017). ALA6, a P4-type ATPase, Is Involved in Heat Stress Responses in Arabidopsis thaliana. *Frontiers in Plant Science*, **8**, 1732.

Oberschall, A., Deák, M., Török, K., Sass, L., Vass, I., Kovács, I., Fehér, A., Dudits, D., and Horváth, G. V. (2000). A novel aldose/aldehyde reductase protects transgenic

plants against lipid peroxidation under chemical and drought stresses. *The Plant Journal*, **24**(4), 437–446.

Oda-Yamamizo, C., Mitsuda, N., Sakamoto, S., Ogawa, D., Ohme-Takagi, M., and Ohmiya, A. (2016). The NAC transcription factor ANAC046 is a positive regulator of chlorophyll degradation and senescence in Arabidopsis leaves. *Scientific reports*, **6**, 23609.

Oerke, E.-C. and Dehne, H.-W. (2004). Safeguarding production - losses in major crops and the role of crop protection. *Crop protection*, **23**(4), 275–285.

Ogata, K. (2009). *Modern control engineering*. Pearson, India: Prentice hall, 5th edition.

Olsen, A. N., Ernst, H. A., Leggio, L. L., and Skriver, K. (2005). NAC transcription factors: structurally distinct, functionally diverse. *Trends in Plant Science*, **10**(2), 79–87.

O'Malley, R. C., Huang, S.-s. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A., and Ecker, J. R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, **165**(5), 1280–1292.

Ó'Maoiléidigh, D. S., Graciet, E., and Wellmer, F. (2014). Gene networks controlling Arabidopsis thaliana flower development. *New Phytologist*, **201**(1), 16–30.

Ouwerkerk, P. B. and Meijer, A. H. (2011). Yeast one-hybrid screens for detection of transcription factor DNA interactions. In *Plant Reverse Genetics*, pages 211–227. Springer.

Ow, D. S.-W., Nissom, P. M., Philp, R., Oh, S. K.-W., and Yap, M. G.-S. (2006). Global transcriptional analysis of metabolic burden due to plasmid maintenance in Escherichia coli DH5α during batch fermentation. *Enzyme and Microbial Technology*, **39**(3), 391–398.

Özgür, A., Vu, T., Erkan, G., and Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, **24**(13), i277–i285.

Pagnussat, G. C., Yu, H.-J., Ngo, Q. A., Rajani, S., Mayalagu, S., Johnson, C. S., Capron, A., Xie, L.-F., Ye, D., and Sundaresan, V. (2005). Genetic and molecular identification of genes required for female gametophyte development and function in Arabidopsis. *Development*, **132**(3), 603–614.

Pajoro, A., Biewers, S., Dougali, E., Leal Valentim, F., Mendes, M. A., Porri, A., Coupland, G., Van de Peer, Y., Van Dijk, A. D., Colombo, L., *et al.* (2014). The (r) evolution of gene regulatory networks controlling Arabidopsis plant reproduction: a two-decade history. *Journal of Experimental Botany*, **65**(17), 4731–4745.

Parent, B., Turc, O., Gibon, Y., Stitt, M., and Tardieu, F. (2010). Modelling temperature-compensated physiological rates, based on the co-ordination of responses to temperature of developmental processes. *Journal of Experimental Botany*, **61**(8), 2057–2069.

Parmar, N., Singh, K. H., Sharma, D., Singh, L., Kumar, P., Nanjundan, J., Khan, Y. J., Chauhan, D. K., and Thakur, A. K. (2017). Genetic engineering strategies for biotic and abiotic stress tolerance and quality enhancement in horticultural crops: a comprehensive review. *3 Biotech*, **7**(4), 239.

Pathak, R. K., Giri, P., Taj, G., and Kumar, A. (2013). Molecular modeling and docking approach to predict the potential interacting partners involved in various biological processes of MAPK with downstream WRKY transcription factor family in Arabidopsis thaliana. *International Journal of Computational Bioinformatics and In Silico Modeling*, **2**, 262–268.

Patron, N. J., Orzaez, D., Marillonnet, S., Warzecha, H., Matthewman, C., Youles, M., Raitskin, O., Leveau, A., Farré, G., Rogers, C., *et al.* (2015). Standards for plant synthetic biology: a common syntax for exchange of DNA parts. *New Phytologist*, **208**(1), 13–19.

Pauwels, L., Morreel, K., De Witte, E., Lammertyn, F., Van Montagu, M., Boerjan, W., Inzé, D., and Goossens, A. (2008). Mapping methyl jasmonate-mediated transcriptional reprogramming of metabolism and cell cycle progression in cultured Arabidopsis cells. *Proceedings of the National Academy of Sciences*, **105**(4), 1380–1385.

Peltier, J.-B., Ytterberg, A. J., Sun, Q., and van Wijk, K. J. (2004). New functions of the thylakoid membrane proteome of Arabidopsis thaliana revealed by a simple, fast, and versatile fractionation strategy. *Journal of Biological Chemistry*, **279**(47), 49367–49383.

Penfold, C. A. and Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface focus*, **1**(6), 857–870.

Penfold, C. A., Gherman, I., Sybirna, A., and Wild, D. L. (2019). Inferring Gene Regulatory Networks from Multiple Datasets. In *Gene Regulatory Networks*, pages 251–282. Springer.

Penninckx, I. A., Thomma, B. P., Buchala, A., Métraux, J.-P., and Broekaert, W. F. (1998). Concomitant activation of jasmonate and ethylene response pathways is required for induction of a plant defensin gene in Arabidopsis. *The Plant Cell*, **10**(12), 2103–2113.

Pérez-Rodríguez, P., Riano-Pachon, D. M., Corrêa, L. G. G., Rensing, S. A., Kersten, B., and Mueller-Roeber, B. (2009). PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Research*, **38**(suppl_1), D822–D827.

Pieterse, C. M., Leon-Reyes, A., Van der Ent, S., and Van Wees, S. C. (2009). Networking by small-molecule hormones in plant immunity. *Nature chemical biology*, **5**(5), 308.

Poinssot, B., Vandelle, E., Bentéjac, M., Adrian, M., Levis, C., Brygoo, Y., Garin, J., Sicilia, F., Coutos-Thévenot, P., and Pugin, A. (2003). The endopolygalacturonase 1 from Botrytis cinerea activates grapevine defense reactions unrelated to its enzymatic activity. *Molecular Plant-Microbe Interactions*, **16**(6), 553–564.

Pokhilko, A., Hodge, S. K., Stratford, K., Knox, K., Edwards, K. D., Thomson, A. W., Mizuno, T., and Millar, A. J. (2010). Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular Systems Biology*, **6**(1), 416.

Pokhilko, A., Fernández, A. P., Edwards, K. D., Southern, M. M., Halliday, K. J., and Millar, A. J. (2012). The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. *Molecular Systems Biology*, **8**(1), 574.

Polański, K., Gao, B., Mason, S. A., Brown, P., Ott, S., Denby, K. J., and Wild, D. L.

(2017). Bringing numerous methods for expression and promoter analysis to a public cloud computing service. *Bioinformatics*, **34**(5), 884–886.

Ponnala, L., Wang, Y., Sun, Q., and van Wijk, K. J. (2014). Correlation of mRNA and protein abundance in the developing maize leaf. *The Plant Journal*, **78**(3), 424–440.

Povero, G., Loreti, E., Pucciariello, C., Santaniello, A., Di Tommaso, D., Di Tommaso, G., Kapetis, D., Zolezzi, F., Piaggesi, A., and Perata, P. (2011). Transcript profiling of chitosan-treated Arabidopsis seedlings. *Journal of Plant Research*, **124**(5), 619–629.

Pré, M., Atallah, M., Champion, A., De Vos, M., Pieterse, C. M., and Memelink, J. (2008). The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense. *Plant Physiology*, **147**(3), 1347–1357.

Prins, T. W., Tudzynski, P., von Tiedemann, A., Tudzynski, B., Ten Have, A., Hansen, M. E., Tenberge, K., and van Kan, J. A. (2000). Infection strategies of Botrytis cinerea and related necrotrophic pathogens. In *Fungal pathology*, pages 33–64. Springer.

Prud'Homme, B., Gompel, N., Rokas, A., Kassner, V. A., Williams, T. M., Yeh, S.-D., True, J. R., and Carroll, S. B. (2006). Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, **440**(7087), 1050–1053.

Pruneda-Paz, J., Breton, G., Nagel, D., Kang, S., Bonaldi, K., Doherty, C., Ravelo, S., Galli, M., Ecker, J., and Kay, S. (2014). A genome-scale resource for the functional characterization of Arabidopsis transcription factors. *Cell Reports*, **8**(2), 622–632.

Pruneda-Paz, J. L., Breton, G., Para, A., and Kay, S. A. (2009). A Functional Genomics Approach Reveals CHE as a Component of the Arabidopsis Circadian Clock. *Science*, **323**(5920), 1481–1485.

Prusinkiewicz, P. and Lindenmayer, A. (2012). *The algorithmic beauty of plants*. Springer Science & Business Media.

Puranik, S., Sahu, P. P., Srivastava, P. S., and Prasad, M. (2012). NAC proteins: regulation and role in stress tolerance. *Trends in Plant Science*, **17**(6), 369–381.

Purnick, P. E. and Weiss, R. (2009). The second wave of synthetic biology: from modules to systems. *Nature reviews Molecular Cell Biology*, **10**(6), 410–422.

Qi, T., Song, S., Ren, Q., Wu, D., Huang, H., Chen, Y., Fan, M., Peng, W., Ren, C., and Xie, D. (2011). The Jasmonate-ZIM-domain proteins interact with the WD-Repeat/bHLH/MYB complexes to regulate Jasmonate-mediated anthocyanin accumulation and trichome initiation in Arabidopsis thaliana. *The Plant Cell*, **23**(5), 1795–1814.

Qian, Y., Huang, H.-H., Jiménez, J. I., and Del Vecchio, D. (2017). Resource competition shapes the response of genetic circuits. *ACS Synthetic Biology*, **6**(7), 1263–1272.

Quackenbush, J. (2001). Computational genetics: computational analysis of microarray data. *Nature Reviews Genetics*, **2**(6), 418.

Raghavendra, A. S. and Murata, Y. (2017). Signal transduction in stomatal guard cells. *Frontiers in Plant Science*, **8**, 114.

Rama Devi, S., Chen, X., Oliver, D. J., and Xiang, C. (2006). A novel high-throughput genetic screen for stress-responsive mutants of Arabidopsis thaliana reveals new loci involving stress responses. *The Plant Journal*, **47**(4), 652–663.

Ramírez, V., Coego, A., López, A., Agorio, A., Flors, V., and Vera, P. (2009). Drought tolerance in Arabidopsis is controlled by the OCP3 disease resistance regulator. *The Plant Journal*, **58**(4), 578–591.

Ramonell, K., Berrocal-Lobo, M., Koh, S., Wan, J., Edwards, H., Stacey, G., and Somerville, S. (2005). Loss-of-function mutations in chitin responsive genes show increased susceptibility to the powdery mildew pathogen Erysiphe cichoracearum. *Plant Physiology*, **138**(2), 1027–1036.

Ranf, S., Eschen-Lippold, L., Pecher, P., Lee, J., and Scheel, D. (2011). Interplay between calcium signalling and early signalling elements during defence responses to microbe-or damage-associated molecular patterns. *The Plant Journal*, **68**(1), 100–113.

Ransbotyn, V., Yeger-Lotem, E., Basha, O., Acuna, T., Verduyn, C., Gordon, M., Chalifa-Caspi, V., Hannah, M. A., and Barak, S. (2015). A combination of gene expression ranking and co-expression network analysis increases discovery rate in large-scale mutant screens for novel Arabidopsis thaliana abiotic stress genes. *Plant Biotechnology Journal*, **13**(4), 501–513.

Rao, C. V., Wolf, D. M., and Arkin, A. P. (2002). Control, exploitation and tolerance of intracellular noise. *Nature*, **420**(6912), 231.

Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. In *Advanced Lectures on Machine Learning*, pages 63–71. Springer, Berlin, Heidelberg.

Rasmussen, C. E. and Nickisch, H. (2010). Gaussian Processes for Machine Learning (GPML) Toolbox. *Journal of Machine Learning Research*, **11**(Nov), 3011–3015.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Process for Machine Learning*. MIT press.

Rauf, M., Arif, M., Fisahn, J., Xue, G.-P., Balazadeh, S., and Mueller-Roeber, B. (2013). NAC transcription factor SPEEDY HYPONASTIC GROWTH regulates flooding-induced leaf movement in Arabidopsis. *The Plant Cell*, **25**(12), 4941–4955.

Ravasz, E. and Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E*, **67**(2), 026112.

Reece-Hoyes, J. S. and Walhout, A. M. (2012). Yeast one-hybrid assays: a historical and technical perspective. *Methods*, **57**(4), 441–447.

REIS, H., PFIFFI, S., and HAHN, M. (2005). Molecular and functional characterization of a secreted lipase from Botrytis cinerea. *Molecular Plant pathology*, **6**(3), 257–267.

Rhodius, V. A., Segall-Shapiro, T. H., Sharon, B. D., Ghodasara, A., Orlova, E., Tabakh, H., Burkhardt, D. H., Clancy, K., Peterson, T. C., Gross, C. A., *et al.* (2013). Design of orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and promoters. *Molecular systems biology*, **9**(1), 702.

Rinaudo, M. (2006). Chitin and chitosan: properties and applications. *Progress in Polymer Science*, **31**(7), 603–632.

Rizhsky, L., Liang, H., Shuman, J., Shulaev, V., Davletova, S., and Mittler, R. (2004). When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. *Plant Physiology*, **134**(4), 1683–1696.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**(3), R25.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Roth, P. C., Arnold, D. C., and Miller, B. P. (2003). MRNet: A software-based multicast/reduction network for scalable tools. In *Proceedings of the 2003 ACM/IEEE conference on Supercomputing*, page 21. ACM.

Roxas, V. P., Smith Jr, R. K., Allen, E. R., and Allen, R. D. (1997). Overexpression of glutathione S-transferase/glutathioneperoxidase enhances the growth of transgenic tobacco seedlings during stress. *Nature Biotechnology*, **15**(10), 988.

Rushton, P. J., Somssich, I. E., Ringler, P., and Shen, Q. J. (2010). WRKY transcription factors. *Trends in Plant Science*, **15**(5), 247–258.

Rzhetsky, A. and Gomez, S. M. (2001). Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics*, **17**(10), 988–996.

Saint Savage, N., Walker, T., Wieckowski, Y., Schiefelbein, J., Dolan, L., and Monk, N. A. (2008). A mutual support mechanism through intercellular movement of CAPRICE and GLABRA3 can pattern the Arabidopsis root epidermis. *PLoS Biology*, **6**(9), e235.

Sakamoto, H., Maruyama, K., Sakuma, Y., Meshi, T., Iwabuchi, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2004). Arabidopsis Cys2/His2-type zinc-finger proteins function as transcription repressors under drought, cold, and high-salinity stress conditions. *Plant Physiology*, **136**(1), 2734–2746.

Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muniz-Rascado, L., García-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., *et al.* (2012). RegulonDB v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, **41**(D1), D203–D213.

Savin, R. and Nicolas, M. E. (1996). Effects of short periods of drought and high temperature on grain growth and starch accumulation of two malting barley cultivars. *Functional Plant Biology*, **23**(2), 201–210.

Sbarbaro, D. and Murray-Smith, R. (2005). Self-tuning control of non-linear systems using Gaussian process prior models. In *Switching and Learning in Feedback Systems*, pages 140–157. Springer.

Scardoni, G., Tosadori, G., Pratap, S., Spoto, F., and Laudanna, C. (2015). Finding the shortest path with PesCa: a tool for network reconstruction. *F1000Research*, **4**.

Schaefer, R. J., Michno, J.-M., and Myers, C. L. (2017). Unraveling gene function in agricultural species using gene co-expression networks. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1860**(1), 53–63.

Schäfer, J. and Strimmer, K. (2004). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**(6), 754–764.

Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**(16), 2263–2270.

Schaumberg, K. A., Antunes, M. S., Kassaw, T. K., Xu, W., Zalewski, C. S., Medford, J. I., and Prasad, A. (2015). Quantitative characterization of genetic parts and circuits for plant synthetic biology. *Nature Methods*, **13**(1), 94.

Schenk, P. M., Kazan, K., Wilson, I., Anderson, J. P., Richmond, T., Somerville, S. C., and Manners, J. M. (2000). Coordinated plant defense responses in Arabidopsis revealed by microarray analysis. *Proceedings of the National Academy of Sciences*, **97**(21), 11655–11660.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., *et al.* (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, **9**(7), 676.

Schmidhuber, J. and Tubiello, F. N. (2007). Global food security under climate change. *Proceedings of the National Academy of Sciences*, **104**(50), 19703–19708.

Schmidt, M. W., Houseman, A., Ivanov, A. R., and Wolf, D. A. (2007). Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe. *Molecular Systems Biology*, **3**(1), 79.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., and Ragg, T. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, **7**(1), 3.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., *et al.* (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, **473**(7347), 337.

Seelig, G., Soloveichik, D., Zhang, D. Y., and Winfree, E. (2006). Enzyme-free nucleic acid logic circuits. *Science*, **314**(5805), 1585–1588.

Segall-Shapiro, T. H., Sontag, E. D., and Voigt, C. A. (2018). Engineered promoters enable constant gene expression at any copy number in bacteria. *Nature Biotechnology*, **36**(4), 352.

Segarra, G., Van der Ent, S., Trillas, I., and Pieterse, C. (2009). MYB72, a node of convergence in induced systemic resistance triggered by a fungal and a bacterial beneficial microbe. *Plant Biology*, **11**(1), 90–96.

Seo, P. J. and Park, C.-M. (2010). MYB96-mediated abscisic acid signals induce pathogen resistance response by promoting salicylic acid biosynthesis in Arabidopsis. *New Phytologist*, **186**(2), 471–483.

Sessions, A., Burke, E., Presting, G., Aux, G., McElver, J., Patton, D., Dietrich, B., Ho, P., Bacwaden, J., Ko, C., *et al.* (2002). A high-throughput Arabidopsis reverse genetics system. *The Plant Cell*, **14**(12), 2985–2994.

Sewelam, N., Oshima, Y., Mitsuda, N., and Ohme-Takagi, M. (2014). A step towards understanding plant responses to multiple environmental stresses: a genome-wide study. *Plant, Cell & Environment*, **37**(9), 2024–2035.

Shachrai, I., Zaslaver, A., Alon, U., and Dekel, E. (2010). Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. *Molecular Cell*, **38**(5), 758–767.

Shadle, G., Chen, F., Reddy, M. S., Jackson, L., Nakashima, J., and Dixon, R. A. (2007). Down-regulation of hydroxycinnamoyl CoA: shikimate hydroxycinnamoyl transferase in transgenic alfalfa affects lignification, development and forage quality. *Phytochemistry*, **68**(11), 1521–1529.

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J. J., Qiu, J.-L., *et al.* (2013). Targeted genome modification of crop plants using a CRISPR-Cas system. *Nature Biotechnology*, **31**(8), 686.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**(11), 2498–2504.

Sharma, R., De Vleesschauwer, D., Sharma, M. K., and Ronald, P. C. (2013). Recent advances in dissecting stress-regulatory crosstalk in rice. *Molecular Plant*, **6**(2), 250–260.

Shen, Q.-H., Saijo, Y., Mauch, S., Biskup, C., Bieri, S., Keller, B., Seki, H., Ülker, B., Somssich, I. E., and Schulze-Lefert, P. (2007). Nuclear activity of MLA immune receptors links isolate-specific and basal disease-resistance responses. *Science*, **315**(5815), 1098–1103.

Shepard, J. F. and Totten, R. E. (1977). Mesophyll cell protoplasts of potato: isolation, proliferation, and plant regeneration. *Plant Physiology*, **60**(2), 313–316.

Shikata, M., Matsuda, Y., Ando, K., Nishii, A., Takemura, M., Yokota, A., and Kohchi, T. (2004). Characterization of Arabidopsis ZIM, a member of a novel plant-specific GATA factor gene family. *Journal of Experimental Botany*, **55**(397), 631–639.

Shiu, S.-H., Shih, M.-C., and Li, W.-H. (2005). Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiology*, **139**(1), 18–26.

Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**(2), 261–274.

Siddiqui, M. S., Thodey, K., Trenchard, I., and Smolke, C. D. (2012). Advancing secondary metabolite biosynthesis in yeast with synthetic biology tools. *FEMS Yeast Research*, **12**(2), 144–170.

Singh, K. B., Foley, R. C., and Oñate-Sánchez, L. (2002). Transcription factors in plant defense and stress responses. *Current Opinion in Plant Biology*, **5**(5), 430–436.

Snelson, E., Ghahramani, Z., and Rasmussen, C. E. (2004). Warped gaussian processes. In *Advances in neural information processing systems*, pages 337–344.

Solano, R., Stepanova, A., Chao, Q., and Ecker, J. R. (1998). Nuclear events in ethylene signaling: a transcriptional cascade mediated by ETHYLENE-INSENSITIVE3 and ETHYLENE-RESPONSE-FACTOR1. *Genes & Development*, **12**(23), 3703–3714.

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, **4**.

Sparks, E. E. and Benfey, P. N. (2017). Tissue-Specific Transcriptome Profiling in Arabidopsis Roots. In *Plant Genomics*, pages 107–122. Springer, Humana Press, New York, NY.

Spinelli, S. V., Martin, A. P., Viola, I. L., Gonzalez, D. H., and Palatnik, J. F. (2011). A mechanistic link between STM and CUC1 during Arabidopsis development. *Plant Physiology*, pages pp–111.

Spoel, S. H., Koornneef, A., Claessens, S. M., Korzelius, J. P., Van Pelt, J. A., Mueller, M. J., Buchala, A. J., Métraux, J.-P., Brown, R., Kazan, K., *et al.* (2003). NPR1 modulates cross-talk between salicylate-and jasmonate-dependent defense pathways through a novel function in the cytosol. *The Plant Cell*, **15**(3), 760–770.

Sreevidya, V., Rao, C. S., Sullia, S., Ladha, J. K., and Reddy, P. M. (2006). Metabolic engineering of rice with soybean isoflavone synthase for promoting nodulation gene expression in rhizobia. *Journal of Experimental Botany*, **57**(9), 1957–1969.

Staiger, D. and Brown, J. W. (2013). Alternative splicing at the intersection of biological timing, development, and stress responses. *The Plant Cell*, **25**(10), 3640–3656.

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., and Borgwardt, K. M. (2010). A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology*, **17**(3), 355–367.

Steinacher, A., Bates, D. G., Akman, O. E., and Soyer, O. S. (2016). Nonlinear dynamics in gene regulation promote robustness and evolvability of gene expression levels. *PLoS One*, **11**(4), e0153295.

Strange, R. N. and Scott, P. R. (2005). Plant disease: a threat to global food security. *Annual Review of Phytopathology*, **43**, 83–116.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**(5643), 249–255.

Studham, M. E., Tjärnberg, A., Nordling, T. E., Nelander, S., and Sonnhammer, E. L. (2014). Functional association networks as priors for gene regulatory network inference. *Bioinformatics*, **30**(12), i130–i138.

Stumpf, M. P., Wiuf, C., and May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences*, **102**(12), 4221–4224.

Suzuki, N., Rivero, R. M., Shulaev, V., Blumwald, E., and Mittler, R. (2014). Abiotic and biotic stress combinations. *New Phytologist*, **203**(1), 32–43.

Swift, J. and Coruzzi, G. M. (2017). A matter of time - How transient transcription factor interactions create dynamic gene regulatory networks. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1860**(1), 75–83.

Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T., *et al.* (2000). Sequence and analysis of chromosome 5 of the plant Arabidopsis thaliana. *Nature*, **408**(6814), 823–826.

Takada, S., Takada, N., and Yoshida, A. (2013). ATML1 promotes epidermal cell differentiation in Arabidopsis shoots. *Development*, **140**(9), 1919–1923.

Team, R. C. *et al.* (2013). R: A language and environment for statistical computing.

Teige, M., Scheikl, E., Eulgem, T., Dóczi, R., Ichimura, K., Shinozaki, K., Dangl, J. L., and Hirt, H. (2004). The MKK2 pathway mediates cold and salt stress signaling in Arabidopsis. *Molecular Cell*, **15**(1), 141–152.

Thaler, J. S. and Bostock, R. M. (2004). Interactions between abscisic-acid-mediated responses and plant resistance to pathogens and insects. *Ecology*, **85**(1), 48–58.

Thomma, B. P., Eggermont, K., Penninckx, I. A., Mauch-Mani, B., Vogelsang, R., Cammue, B. P., and Broekaert, W. F. (1998). Separate jasmonate-dependent and salicylate-dependent defense-response pathways in Arabidopsis are essential for resistance to distinct microbial pathogens. *Proceedings of the National Academy of Sciences*, **95**(25), 15107–15111.

Thomson, A. R., Wood, C. W., Burton, A. J., Bartlett, G. J., Sessions, R. B., Brady, R. L., and Woolfson, D. N. (2014). Computational design of water-soluble α-helical barrels. *Science*, **346**(6208), 485–488.

Tian, Q., Stepaniants, S. B., Mao, M., Weng, L., Feetham, M. C., Doyle, M. J., Eugene, C. Y., Dai, H., Thorsson, V., Eng, J., *et al.* (2004). Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Molecular & Cellular Proteomics*, **3**(10), 960–969.

Tian, T. and Burrage, K. (2006). Stochastic models for regulatory networks of the genetic toggle switch. *Proceedings of the National Academy of Sciences*, **103**(22), 8372–8377.

Tian, T., Burrage, K., Burrage, P. M., and Carletti, M. (2007). Stochastic delay differential equations for genetic regulatory networks. *Journal of Computational and Applied Mathematics*, **205**(2), 696–707.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tilman, D., Balzer, C., Hill, J., and Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*, **108**(50), 20260–20264.

Toettcher, J. E., Gong, D., Lim, W. A., and Weiner, O. D. (2011). Light-based feedback for controlling intracellular signaling dynamics. *Nature methods*, **8**(10), 837.

Toledo-Ortiz, G., Huq, E., and Quail, P. H. (2003). The Arabidopsis basic/helix-loop-helix transcription factor family. *The Plant Cell*, **15**(8), 1749–1770.

Ton, J. and Mauch-Mani, B. (2004). β-amino-butyric acid-induced resistance against necrotrophic pathogens is based on ABA-dependent priming for callose. *The Plant Journal*, **38**(1), 119–130.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**(31), 187–202.

Torres, M. A. (2010). ROS in biotic interactions. *Physiologia Plantarum*, **138**(4), 414–429.

Torres, M. A., Jones, J. D., and Dangl, J. L. (2006). Reactive oxygen species signaling in response to pathogens. *Plant Physiology*, **141**(2), 373–378.

Tran, L.-S. P., Nakashima, K., Sakuma, Y., Simpson, S. D., Fujita, Y., Maruyama, K., Fujita, M., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2004). Isolation

and functional analysis of Arabidopsis stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. *The Plant Cell*, **16**(9), 2481–2498.

Trigg, S. A., Garza, R. M., MacWilliams, A., Nery, J. R., Bartlett, A., Castanon, R., Goubil, A., Feeney, J., O'Malley, R., Shao-shan, C. H., *et al.* (2017). CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. *Nature Methods*, **14**(8), 819.

Tscharntke, T., Clough, Y., Wanger, T. C., Jackson, L., Motzke, I., Perfecto, I., Vandermeer, J., and Whitbread, A. (2012). Global food security, biodiversity conservation and the future of agricultural intensification. *Biological Conservation*, **151**(1), 53–59.

Tsuda, K. and Katagiri, F. (2010). Comparing signaling mechanisms engaged in pattern-triggered and effector-triggered immunity. *Current Opinion in Plant Biology*, **13**(4), 459–465.

Tsuda, K., Sato, M., Stoddard, T., Glazebrook, J., and Katagiri, F. (2009). Network properties of robust immunity in plants. *PLoS Genetics*, **5**(12), e1000772.

Tsuda, K., Qi, Y., Nguyen, L. V., Bethke, G., Tsuda, Y., Glazebrook, J., and Katagiri, F. (2012). An efficient Agrobacterium-mediated transient transformation of Arabidopsis. *The Plant Journal*, **69**(4), 713–719.

Tubiello, F. N., Salvatore, M., Ferrara, A. F., House, J., Federici, S., Rossi, S., Biancalani, R., Condor Golec, R. D., Jacobs, H., Flammini, A., *et al.* (2015). The contribution of agriculture, forestry and other land use activities to global warming, 1990–2012. *Global Change Biology*, **21**(7), 2655–2660.

Turner, R., Rabalais, N., Justic, D., and Dortch, Q. (2003). Global patterns of dissolved N, P and Si in large rivers. *Biogeochemistry*, **64**(3), 297–317.

Twycross, J., Band, L. R., Bennett, M. J., King, J. R., and Krasnogor, N. (2010). Stochastic and deterministic multiscale models for systems biology: an auxin-transport case study. *BMC Systems Biology*, **4**(1), 34.

Umezawa, T., Fujita, M., Fujita, Y., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2006). Engineering drought tolerance in plants: discovering and tailoring genes to unlock the future. *Current Opinion in Biotechnology*, **17**(2), 113–122.

Underwood, W., Zhang, S., and He, S. Y. (2007). The Pseudomonas syringae type III effector tyrosine phosphatase HopAO1 suppresses innate immunity in Arabidopsis thaliana. *The Plant Journal*, **52**(4), 658–672.

Uno, Y., Furihata, T., Abe, H., Yoshida, R., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2000). Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proceedings of the National Academy of Sciences*, **97**(21), 11632–11637.

Vailleau, F., Daniel, X., Tronchet, M., Montillet, J.-L., Triantaphylides, C., and Roby, D. (2002). A R2R3-MYB gene, AtMYB30, acts as a positive regulator of the hypersensitive cell death program in plants in response to pathogen attack. *Proceedings of the National Academy of Sciences*, **99**(15), 10179–10184.

Van Kan, J., Van't Klooster, J., Wagemakers, C., Dees, D., and Van der Vlugt-Bergmans, C. (1997). Cutinase A of Botrytis cinerea is expressed, but not essen-

tial, during penetration of gerbera and tomato. *Molecular Plant-Microbe Interactions*, **10**(1), 30–38.

van Kan, J. A. (2006). Licensed to kill: the lifestyle of a necrotrophic plant pathogen. *Trends in Plant Science*, **11**(5), 247–253.

van Mourik, S., van Dijk, A. D., de Gee, M., Immink, R. G., Kaufmann, K., Angenent, G. C., van Ham, R. C., and Molenaar, J. (2010). Continuous-time modeling of cell fate determination in Arabidopsis flowers. *BMC Systems Biology*, **4**(1), 101.

Veronese, P., Nakagami, H., Bluhm, B., AbuQamar, S., Chen, X., Salmeron, J., Dietrich, R. A., Hirt, H., and Mengiste, T. (2006). The membrane-anchored BOTRYTIS-INDUCED KINASE1 plays distinct roles in Arabidopsis resistance to necrotrophic and biotrophic pathogens. *The Plant Cell*, **18**(1), 257–273.

Verpoorte, R. and Alfermann, A. W. (2013). *Metabolic engineering of plant secondary metabolism*. Springer Science & Business Media, Wiley Online Library.

Vidyasagar, M. (1998). Statistical learning theory and randomized algorithms for control. *IEEE Control Systems*, **18**(6), 69–85.

Vinayagam, A., Gibson, T. E., Lee, H.-J., Yilmazel, B., Roesel, C., Hu, Y., Kwon, Y., Sharma, A., Liu, Y.-Y., Perrimon, N., and Barabási, A.-L. (2016). Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Sciences*, **113**(18), 4976–4981.

Vinh, N. X., Chetty, M., Coppel, R., and Wangikar, P. P. (2012). Issues impacting genetic network reverse engineering algorithm validation using small networks. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, **1824**(12), 1434–1441.

Voelker, S. L., Lachenbruch, B., Meinzer, F. C., Jourdes, M., Ki, C., Patten, A. M., Davin, L. B., Lewis, N. G., Tuskan, G. A., Gunter, L., *et al.* (2010). Antisense down-regulation of 4CL expression alters lignification, tree growth and saccharification potential of field-grown poplar. *Plant Physiology*, pages pp–110.

Wan, J., Zhang, X.-C., Neece, D., Ramonell, K. M., Clough, S., Kim, S.-y., Stacey, M. G., and Stacey, G. (2008). A LysM receptor-like kinase plays a critical role in chitin signaling and fungal resistance in Arabidopsis. *The Plant Cell*, **20**(2), 471–481.

Wang, D., Amornsiripanitch, N., and Dong, X. (2006). A genomic approach to identify regulatory nodes in the transcriptional network of systemic acquired resistance in plants. *PLoS Pathogens*, **2**(11), e123.

Wang, D., Pajerowska-Mukhtar, K., Culler, A. H., and Dong, X. (2007). Salicylic acid inhibits pathogen growth in plants through repression of the auxin signaling pathway. *Current Biology*, **17**(20), 1784–1790.

Wang, D., Eraslan, B., Wieland, T., Hallstrom, B. M., Hopf, T., Zolg, D. P., Zecha, J., Asplund, A., Li, L.-h., Meng, C., *et al.* (2018). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *bioRxiv*, page 357137.

Wang, H., Xu, Q., Kong, Y.-H., Chen, Y., Duan, J.-Y., Wu, W.-H., and Chen, Y.-F. (2014). Arabidopsis WRKY45 transcription factor activates PHT1; 1 expression in response to phosphate starvation. *Plant Physiology*, pages pp–113.

Wang, R.-L., Stec, A., Hey, J., Lukens, L., and Doebley, J. (1999). The limits of selection during maize domestication. *Nature*, **398**(6724), 236.

Wang, Z. and Huang, B. (2004). Physiological recovery of Kentucky bluegrass from simultaneous drought and heat stress. *Crop Science*, **44**(5), 1729–1736.

Weber, C., Nover, L., and Fauth, M. (2008). Plant stress granules and mRNA processing bodies are distinct from heat stress granules. *The Plant Journal*, **56**(4), 517–530.

Wehner, N., Hartmann, L., Ehlert, A., Böttner, S., Oñate-Sánchez, L., and Dröge-Laser, W. (2011). High-throughput protoplast transactivation (PTA) system for the analysis of Arabidopsis transcription factor function. *The Plant Journal*, **68**(3), 560–569.

Weiberg, A., Wang, M., Lin, F.-M., Zhao, H., Zhang, Z., Kaloshian, I., Huang, H.-D., and Jin, H. (2013). Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science*, **342**(6154), 118–123.

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.-M., Pham, G. M., Nicotra, A. B., Gregorio, G. B., Jagadish, S. K., Septiningsih, E. M., Bonneau, R., and Purugganan, M. (2016). EGRINs (Environmental Gene Regulatory Influence Networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *The Plant Cell*, **28**(10), 2365–2384.

Williams, B., Kabbage, M., Kim, H.-J., Britt, R., and Dickman, M. B. (2011). Tipping the balance: Sclerotinia sclerotiorum secreted oxalic acid suppresses host defenses by manipulating the host redox environment. *PLoS Pathogens*, **7**(6), e1002107.

Williamson, B., Tudzynski, B., Tudzynski, P., and Van Kan, J. A. L. (2007). Botrytis cinerea: the cause of grey mould disease. *Molecular Plant Pathology*, **8**(5), 561–580.

Willmann, R., Lajunen, H. M., Erbs, G., Newman, M.-A., Kolb, D., Tsuda, K., Katagiri, F., Fliegmann, J., Bono, J.-J., Cullimore, J. V., *et al.* (2011). Arabidopsis lysin-motif proteins LYM1 LYM3 CERK1 mediate bacterial peptidoglycan sensing and immunity to bacterial infection. *Proceedings of the National Academy of Sciences*, **108**(49), 19824–19829.

Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R., Cooke, E., Jenkins, D. J., Penfold, C. A., Baxter, L., Breeze, E., Kiddle, S. J., Rhodes, J., Atwell, S., Kliebenstein, D. J., Kim, Y.-s., Stegle, O., Borgwardt, K., Zhang, C., Tabrett, A., Legaie, R., Moore, J., Finkenstadt, B., Wild, D. L., Mead, A., Rand, D., Beynon, J., Ott, S., Buchanan-Wollaston, V., and Denby, K. J. (2012). Arabidopsis defense against Botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *The Plant Cell*, **24**(9), 3530–3557.

Windram, O. P., Rodrigues, R. T., Lee, S., Haines, M., and Bayer, T. S. (2017). Engineering microbial phenotypes through rewiring of genetic networks. *Nucleic Acids Research*, **45**(8), 4984–4993.

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of" guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, **6**(1), 1.

Wood, S., Sebastian, K., and Scherr, S. (2000). Pilot analysis of global ecosystems: agroecosystems, a joint study by International Food Policy Research Institute and World Resources Institute. *International Food Policy Research Institute and World Resources Institute, Washington DC*.

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V.,

and Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, **20**(9), 1377–1419.

Wright, C. A. and Beattie, G. A. (2004). Pseudomonas syringae pv. tomato cells encounter inhibitory levels of water stress during the hypersensitive response of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, **101**(9), 3269–3274.

Wu, F.-H., Shen, S.-C., Lee, L.-Y., Lee, S.-H., Chan, M.-T., and Lin, C.-S. (2009a). Tape-Arabidopsis Sandwich-a simpler Arabidopsis protoplast isolation method. *Plant Methods*, **5**(1), 16.

Wu, G., Nie, L., and Zhang, W. (2008a). Integrative analyses of posttranscriptional regulation in the yeast Saccharomyces cerevisiae using transcriptomic and proteomic data. *Current Microbiology*, **57**(1), 18–22.

Wu, H., Kerr, M., Cui, X., and Churchill, G. (2003). MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. *The Analysis of Gene Expression Data*, pages 313–341.

Wu, L. F., Hughes, T. R., Davierwala, A. P., Robinson, M. D., Stoughton, R., and Altschuler, S. J. (2002). Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature Genetics*, **31**(3), 255.

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008b). Network-based global inference of human disease genes. *Molecular Systems Biology*, **4**(1), 189.

Wu, Y., Deng, Z., Lai, J., Zhang, Y., Yang, C., Yin, B., Zhao, Q., Zhang, L., Li, Y., Yang, C., *et al.* (2009b). Dual function of Arabidopsis ATAF1 in abiotic and biotic stress responses. *Cell Research*, **19**(11), 1279.

Xiang, T., Zong, N., Zou, Y., Wu, Y., Zhang, J., Xing, W., Li, Y., Tang, X., Zhu, L., Chai, J., *et al.* (2008). Pseudomonas syringae effector AvrPto blocks innate immunity by targeting receptor kinases. *Current Biology*, **18**(1), 74–80.

Xing, D.-H., Lai, Z.-B., Zheng, Z.-Y., Vinod, K., Fan, B.-F., and Chen, Z.-X. (2008). Stress-and pathogen-induced Arabidopsis WRKY48 is a transcriptional activator that represses plant basal defense. *Molecular Plant*, **1**(3), 459–470.

Xiong, L. and Zhu, J.-K. (2001). Abiotic stress signal transduction in plants: molecular and genetic perspectives. *Physiologia Plantarum*, **112**(2), 152–166.

Xu, G., Greene, G. H., Yoo, H., Liu, L., Marqués, J., Motley, J., and Dong, X. (2017). Global translational reprogramming is a fundamental layer of immune regulation in plants. *Nature*, **545**(7655), 487.

Yamada, Y., Yang, Z.-Q., and Tang, D.-T. (1986). Plant regeneration from protoplast-derived callus of rice (Oryza sativa L.). *Plant Cell Reports*, **5**(2), 85–88.

Yang, F., Mitra, P., Zhang, L., Prak, L., Verhertbruggen, Y., Kim, J.-S., Sun, L., Zheng, K., Tang, K., Auer, M., *et al.* (2013). Engineering secondary cell wall deposition in plants. *Plant Biotechnology Journal*, **11**(3), 325–335.

Yang, Y. and Karlson, D. T. (2011). Overexpression of AtCSP4 affects late stages of embryo development in Arabidopsis. *Journal of Experimental Botany*, **62**(6), 2079–2091.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**(4), e15.

Yant, L., Mathieu, J., Dinh, T. T., Ott, F., Lanz, C., Wollmann, H., Chen, X., and Schmid, M. (2010). Orchestration of the floral transition and floral development in Arabidopsis by the bifunctional transcription factor APETALA2. *The Plant Cell*, **22**(7), 2156–2170.

Yao, C.-W., Hsu, B.-D., and Chen, B.-S. (2011). Constructing gene regulatory networks for long term photosynthetic light acclimation in Arabidopsis thaliana. *BMC Bioinformatics*, **12**(1), 335.

Yasuda, M., Ishikawa, A., Jikumaru, Y., Seki, M., Umezawa, T., Asami, T., Maruyama-Nakashita, A., Kudo, T., Shinozaki, K., Yoshida, S., *et al.* (2008). Antagonistic interaction between systemic acquired resistance and the abscisic acid–mediated abiotic stress response in Arabidopsis. *The Plant Cell*, **20**(6), 1678–1692.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, **13**(1), 134.

Yi, T.-M., Huang, Y., Simon, M. I., and Doyle, J. (2000). Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proceedings of the National Academy of Sciences*, **97**(9), 4649–4653.

Yilmaz, A., Mejia-Guerra, M. K., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). AGRIS: the Arabidopsis gene regulatory information server, an update. *Nucleic Acids Research*, **39**(suppl 1), D1118–D1122.

Yip, K. Y., Alexander, R. P., Yan, K.-K., and Gerstein, M. (2010). Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS One*, **5**(1), e8121.

Yoo, S.-D., Cho, Y.-H., and Sheen, J. (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nature Protocols*, **2**(7), 1565.

Yoshida, T., Fujita, Y., Maruyama, K., Mogami, J., Todaka, D., Shinozaki, K., and YAMAGUCHI-SHINOZAKI, K. (2015). Four A rabidopsis AREB/ABF transcription factors function predominantly in gene expression downstream of SnRK2 kinases in abscisic acid signalling in response to osmotic stress. *Plant, Cell & Environment*, **38**(1), 35–49.

Yu, H. and Gerstein, M. (2006). Genomic analysis of the hierarchical structure of regulatory networks. *Proceedings of the National Academy of Sciences*, **103**(40), 14724–14731.

Yu, T., Wang, L., Yin, Y., Wang, Y., and Zheng, X. (2008). Effect of chitin on the antagonistic activity of Cryptococcus laurentii against Penicillium expansum in pear fruit. *International Journal of Food Microbiology*, **122**(1-2), 44–48.

Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Research*, **37**(suppl_2), W247–W252.

Zarei, A., Körbes, A. P., Younessi, P., Montiel, G., Champion, A., and Memelink, J. (2011). Two GCC boxes and AP2/ERF-domain transcription factor ORA59 in jasmonate/ethylene-mediated activation of the PDF1.2 promoter in Arabidopsis. *Plant Molecular Biology*, **75**(4), 321–331.

Zhang, H., Li, R., and Liu, W. (2011a). Effects of chitin and its derivative chitosan on postharvest decay of fruits: a review. *International Journal of Molecular Sciences*, **12**(2), 917–934.

Zhang, H.-M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H., and Guo, A.-Y. (2011b). AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Research*, **40**(D1), D144–D149.

Zhang, J., Li, W., Xiang, T., Liu, Z., Laluk, K., Ding, X., Zou, Y., Gao, M., Zhang, X., Chen, S., *et al.* (2010). Receptor-like cytoplasmic kinases integrate signaling from multiple plant immune receptors and are targeted by a Pseudomonas syringae effector. *Cell Host & Microbe*, **7**(4), 290–301.

Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E., and Hogenesch, J. B. (2014). A circadian gene expression atlas in mammals: implications for biology and medicine. *Proceedings of the National Academy of Sciences*, **111**(45), 16219–16224.

Zheng, Z., Qamar, S. A., Chen, Z., and Mengiste, T. (2006). Arabidopsis WRKY33 transcription factor is required for resistance to necrotrophic fungal pathogens. *The Plant Journal*, **48**(4), 592–605.

Zhu, X.-G., de Sturler, E., and Long, S. P. (2007). Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate: a numerical simulation using an evolutionary algorithm. *Plant Physiology*, **145**(2), 513–526.

Zhu, Z., An, F., Feng, Y., Li, P., Xue, L., Mu, A., Jiang, Z., Kim, J.-M., To, T. K., Li, W., *et al.* (2011). Derepression of ethylene-stabilized transcription factors (EIN3/EIL1) mediates jasmonate and ethylene signaling synergy in Arabidopsis. *Proceedings of the National Academy of Sciences*, **108**(30), 12539–12544.

Zipfel, C. (2008). Pattern-recognition receptors in plant innate immunity. *Current Opinion in Immunology*, **20**(1), 10–16.

Zipfel, C., Robatzek, S., Navarro, L., Oakeley, E. J., Jones, J. D., Felix, G., and Boller, T. (2004). Bacterial disease resistance in Arabidopsis through flagellin perception. *Nature*, **428**(6984), 764.